UNIVERSITY OF LONDON THESIS

Degree	Year	Name of Author
PhD	2005	CLARE, DOMINIC F.

This thesis comes within category D.

Computational Studies of Protein–Peptide Interactions

Dominic F. Clare

University College
London

A thesis submitted for the

Department of Chemistry

for the degree of Doctor of Philosophy

2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To Mum and Dad. Thanks for everything.

Acknowledgements

I would like to thank my supervisor, Prof. David Clary, for his invaluable help and guidance during this work.

I would also like to thank the now-departed members of the Clary group at UCL, in particular Anthony Meijer, Shervin Moghaddam and Tommy Miller for many helpful discussions and assistance in the laboratory.

I thank my family, in particular my Dad, for moral (not to mention financial) support throughout my years at UCL. I also thank Natalie for her help with proof-reading and, more importantly, for her love, support and encouragement over the past two years that have given me the impetus required to finish this work. Hopefully, I can return the favour soon.

Abstract

The interactions between proteins and peptides in aqueous solution have been investigated using a classical molecular dynamics procedure with a molecular mechanical representation of the potential energy surface. During post-processing of the trajectory an implicit solvation method has been applied in order to calculate the free energy of each of the complex, protein and peptide structures in solution, allowing the binding free energy of the protein–peptide complex to be evaluated. Entropic contributions have been estimated using classical ideal gas thermodynamics. A program has also been developed that systematically mutates each of the peptide residues to alanine and determines the effect on the binding free energy. The method has been applied to the interaction between the oncoprotein Mdm2 and the tumour suppressor peptide p53 and reasonable agreement has been found with previous theoretical and experimental studies. The method has also been extended to the interactions between IQN17, an engineered protein that represents a potential drug target in the HIV-1 gp41-mediated cellular fusion process, and several peptides that have been shown to inhibit cellular fusion. The key residues in the binding of each protein–peptide system have been identified and quantitative information regarding the factors influencing binding obtained and compared with available experimental data, demonstrating encouraging agreement with analogous alanine scanning experiments.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Biomolecular Systems

1.1 Proteins and Polypeptides

1.1.1 Structure and Function

Proteins are biologically important molecules that consist of one or more polypeptide chains, which are linear heteropolymers of monomeric units called amino acids.[1] There are 20 amino acids that occur naturally and the codes used to identify each amino acid in a polypeptide chain conveniently are given in Appendix A. Proteins found in nature only contain amino acids with L-chirality and a schematic diagram of an amino acid is given in Figure 1.1. In aqueous solution, the amine group becomes protonated, NH_3^+, and the carboxylic acid group is deprotonated, COO^-, resulting in a zwitterionic amino acid form. The central carbon atom of an amino acid, illustrated in green in Figure 1.1, is called the α-carbon and the carbon atoms of the side chain are referred to as the β-, γ-,

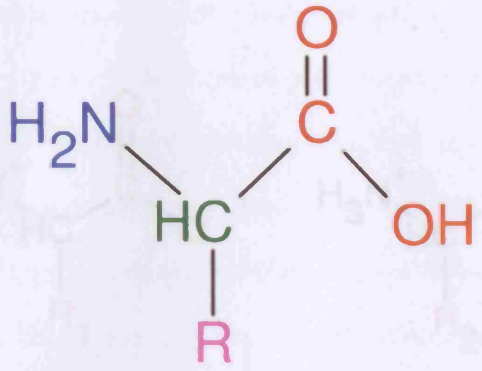

Figure 1.1: Schematic diagram of an amino acid. The amino group is shown in blue, the carboxylic acid group is shown in red and the α-carbon is shown in green.

δ-carbon and so on. Amino acids are distinguished by their side chain groups, which can range from a hydrogen atom, as in the achiral amino acid glycine, to much bulkier side chains such as the indole ring of tryptophan. For a chain of N amino acids there are 20^N possible amino acid sequences and although only a relatively small percentage of these are viable polypeptides, there is a great diversity in observed protein structure and function.

Two amino acids can join together, as shown in Figure 1.2, by the elimination of a water molecule. The carbon–nitrogen bond formed during this process is called a peptide bond and a chain of amino acids joined together in this way is called a polypeptide. The polypeptide chain has an amino- or N-terminus at one end of the chain carrying the -NH_3^+ group, and a carboxy- or C-terminus at the other end of the chain carrying the -COO^- group. By convention, polypeptide chains and amino acid sequences are given with the N-terminus to the left and

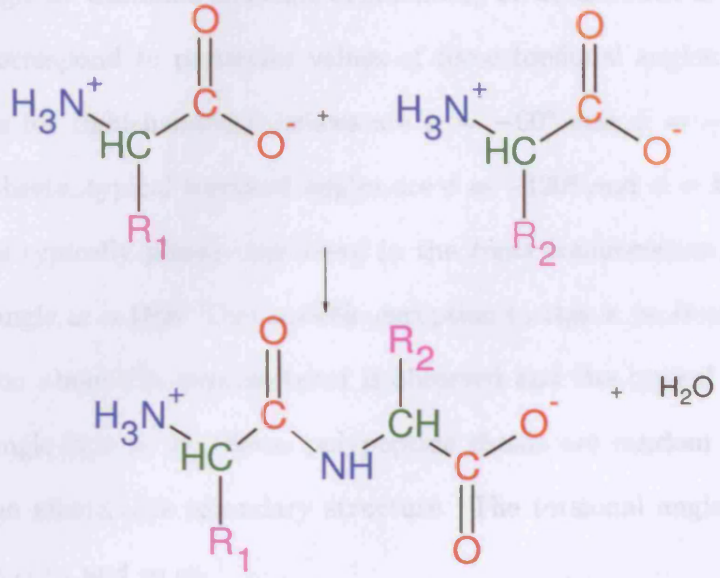

Figure 1.2: Schematic diagram of peptide bond formation. The C–N bond formed is termed the peptide bond.

the C-terminus to the right.

The primary structure of a polypeptide chain refers to the sequence of amino acids in the chain. The local folding of a polypeptide chain is referred to as its secondary structure and the way that these elements of secondary structure are arranged in space is described as the tertiary structure. The quaternary structure describes how several subunits (connected or unconnected polypeptide chains) fold and join together to form the complete functional protein. The secondary structure can be described in terms of the torsional angles along the peptide backbone, where torsions about the N–C_α bond are described in terms of

the dihedral angle ϕ, torsions about the C_α–C bond are described in terms of the dihedral angle ψ and torsions about the C–N peptide bond are defined in terms of the dihedral angle ω. Common elements of secondary structure such as α-helices and β-sheets correspond to particular values of these torsional angles. Typical torsional angles for right-handed α-helices are $\phi = -60°$ and $\psi = -60°$. For antiparallel β-sheets, typical torsional angles are $\phi = -120°$ and $\psi = 120°$. The peptide bond is typically planar and found in the *trans* conformation resulting in a torsional angle $\omega = 180°$. One notable exception to this is proline, where a *cis* conformation about the peptide bond is observed and the typical value for the torsional angle is $\omega = 0°$. Some polypeptide chains are random coils and have little or no identifiable secondary structure. The torsional angles of each side chain are χ_1, χ_2 and so on.[1]

Protein structures can be determined in atomic detail using techniques such as X-ray and neutron diffraction for crystallised protein structures. Proteins can also be studied in solution using nuclear magnetic resonance (NMR) spectroscopy. The atomic structures of a vast number of proteins, approximately 29000 at present, are available from resources such as the Protein Data Bank (PDB).[2]

The biological activity of vital proteins is intrinsically related to their structure. Enzymes are catalysts that control chemical reactions taking place in biological environments. These reactions can be simple processes such as the hydration of carbon dioxide or more complex transformations of large molecules. Proteins such as haemoglobin, which stores oxygen in red blood cells, also have a key function in storing ions and small molecules within the body and transporting them from one place to another. Hormones such as insulin, which regulates the concentration of

sugars in the blood, play vital roles in coordinating biological processes between different tissues and organs. Antibodies are proteins that bind to foreign bodies, including viruses and bacteria, serving a vital biological function.

The non-covalent interactions between biologically important macromolceules such as proteins, polypeptides and nucleic acids provide the key to understanding many physiological processes including enzyme catalysis, protein and nucleic acid folding, immune response processes and complexation of biologically active molecules. These interactions are of particular importance in the biomolecular systems considered in this work, which are discussed in more detail in the remainder of this chapter.

1.2 Suppression of Tumour Growth

The interaction between the tumour suppressor protein p53 and the cancer-causing protein or oncoprotein Mdm2 is of vital importance in cancer research.[3] The role of p53 in the cell cycle is to preserve the DNA from alterations caused by infrared (IR), ultraviolet (UV) and γ radiation, as well as other stress stimuli. The response of p53 to these various stimuli is to arrest cellular growth, allowing time for damaged DNA to be repaired.[4] The Mdm2 oncoprotein precludes this vital regulatory function by forming tightly bound complexes with p53, thus preventing p53 from binding to DNA. It has been found that p53 is missing or altered in approximately half of all human tumours and Mdm2 is often found in increased amounts in a variety of cancers.[3]

The crystal structures of human and African clawed frog, *Xenopus Laevis* (*X.*

Laevis), Mdm2 in complex with a stretch of p53 have been determined previously.[5] Experiments carried out on these complexes identified three residues in p53 that are vital to the binding of the complex. These residues, Phe-19, Trp-23 and Leu-26, could not be replaced by any other amino acid without at least a threefold increase in IC_{50}, defined as the inhibitory concentration required to reduce activity by 50 per cent.[6] The residue Leu-22 was also found to be highly selective.[6] These residues extend into the hydrophobic binding site of Mdm2.[5] Theoretical studies have been carried out on the human and *X. Laevis* Mdm2–p53 complexes using a methodology similar to that used in this work.[3] The findings were in agreement with the replacement experiments and are discussed in more detail in Chapter 5. This system was chosen to test the methodology implemented in this work because of its biophysical significance and the availability of experimental and theoretical results for comparison. The amino acid sequences and structure of the Mdm2–p53 complexes used in this work are discussed in more detail in Chapter 4.

1.3 Inhibition of HIV-1 Activity

1.3.1 Introduction

It was estimated that 40 million people would have been infected with the human immunodeficiency virus (HIV) by the end of the twentieth century[7] and there is strong evidence that type 1 HIV (HIV-1) is the cause of acquired immunodeficiency syndrome (AIDS).[8–10] Clearly, it is essential to understand the

processes by which HIV infection takes place in order to allow the production of effective anti-HIV drugs and to suggest possible new targets for such drugs.

The HIV viral cell is approximately spherical, with a diameter of roughly 1000 Å, and has complexes of two proteins, gp120 and gp41, protruding from the surface. These protrusions are the envelope glycoproteins that are involved in the process of viral entry into the cells. The active parts of the virus are contained within the core of the spherical cell and can only act upon the target cell once viral entry has been achieved. Within the viral envelope is a matrix protein, p17, and the viral core or capsid, p24, containing the viral machinery such as the viral RNA and the enzymes protease (PR), reverse transcriptase (RT) and integrase. HIV is a retrovirus, meaning that it uses the RNA in its viral core as a template to make DNA within the host cell. The enzyme PR within the viral core is used to break proteins down into smaller units. The role of RT is to convert RNA into viral DNA, which is then transported into the nucleus of the cell, where the enzyme integrase facilitates the insertion of the viral DNA into the host cell's DNA. The HIV DNA is then known as a provirus and is replicated by the host cell. A schematic diagram of HIV illustrating these elements is given in Figure 1.3.

There are a large number of drugs that have been approved by the United States Food and Drug Administration (FDA) that target the RT and PR enzymes of HIV.[7,12,13] These are two of the key parts of the viral core of HIV and these drugs are designed to reduce the rate of viral replication significantly. The mortality rate for people infected with the virus has also been significantly improved as a result of treatment with these drugs.[14,15]

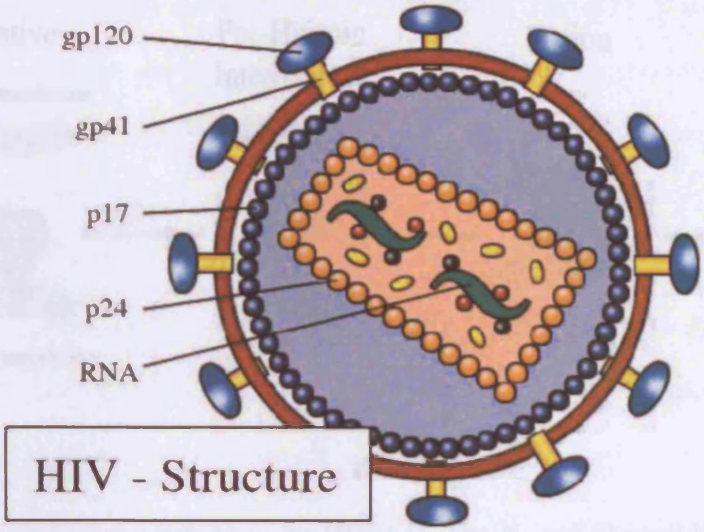

Figure 1.3: Schematic diagram of the HIV cell[11] showing the envelope glycoproteins, gp120 and gp41, as well as the RNA in the core that brings about viral replication.

However, one of the major difficulties in dealing with HIV, and, indeed, one of the reasons that it is such a dangerous pathogen, lies in its ability to mutate rapidly and develop strains that are resistant to treatment with RT- and PR-inhibiting drugs. The rapid emergence of such drug-resistant strains of HIV has become a widespread and growing problem,[16] meaning that new drugs, possibly targeting different aspects of HIV-1 activity, must be developed. Another difficulty in this approach is that these drugs can only ever treat HIV infection rather than preventing it. Furthermore, since these anti-RT and anti-PR drugs are not effective until the virus has fused with the human cell, they must penetrate the human cell, which can lead to debilitating side effects.

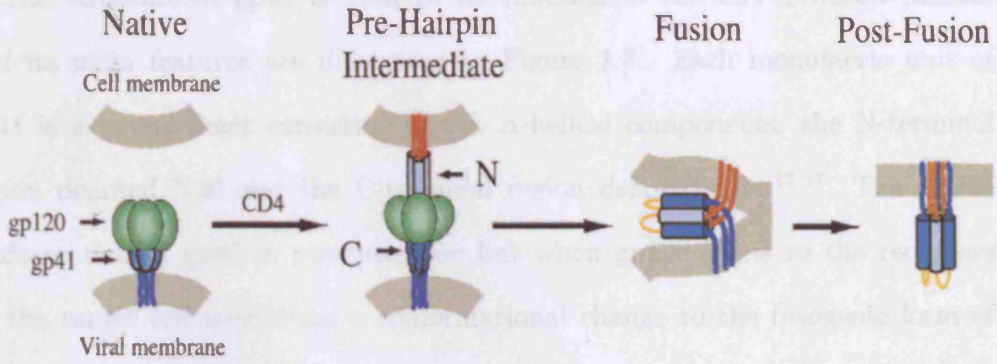

Figure 1.4: Viral entry mechanism for HIV-1.[19] The N- and C-peptides of gp41 are shown in grey and blue, respectively. The conformational change from the pre-hairpin intermediate to the fusion-active form of gp41 brings the two cells together and allows fusion to take place.

1.4 Inhibition of HIV-1 Viral Entry

1.4.1 Mechanism of Viral Entry

HIV-1 is an enveloped virus and its entry into human cells is controlled by the envelope glycoproteins gp120 and gp41, found in a complex on the surface of the viral cell. The mechanism for viral entry involves several steps,[17–19] providing several potential targets for inhibitory drugs,[19–24] and is shown in Figure 1.4. The function of gp120 is to bind to receptors on the target human cell while gp41 mediates the fusion of the viral and target cells. In this work, the main focus is on the function of gp41 and inhibiting the essential conformational change that allows gp41-mediated fusion to take place.

The structure of gp41 is vital to its function in the HIV-1 fusion process and its main features are illustrated in Figure 1.5. Each monomeric unit of gp41 is a heterodimer consisting of two α-helical components; the N-terminal region denoted N36 and the C-terminal region denoted C34.[12,17] The native conformation of gp41 is non-fusogenic but when gp120 binds to the receptors on the target cell membrane a conformational change to the fusogenic form of gp41 takes place. This fusion-active conformation of gp41 consists of a trimer of heterodimers. The three N-peptide helices come together to form a parallel trimeric core with the C-peptide helices bound in an anti-parallel fashion along hydrophobic grooves on the surface of this core. Each of these grooves has a deep hydrophobic pocket at one end.

Once the human and viral cells are attached an intermediate of gp41, termed the pre-hairpin intermediate, is transiently exposed, providing the target for many of the fusion inhibitors discussed here. It is possible to inhibit gp41-mediated fusion by binding molecules to the N-peptide or C-peptide regions of the transiently-exposed intermediate, preventing the essential conformational change. Once the six-helical bundle is formed it is very stable and inhibitors are not effective in the same manner. In the pre-hairpin intermediate species the glycine-rich fusion peptide (denoted fp in Figure 1.5) is exposed and interacts with the target cell membrane. The intermediate then gradually gives way to the fusogenic conformation, bringing the human and viral cells into close proximity and allowing membrane fusion to take place.[12,17,19,25-28]

Many other enveloped viruses undergo similar membrane fusion processes. It is therefore likely that breakthroughs in understanding and inhibiting the HIV-1

fusion process could lead to advances in the treatment of such viruses, including influenza,[29] Ebola,[30] Visna,[31] human respiratory syncytial virus (HRSV)[32] and the Moloney murine leukaemia virus (MMLV).[33]

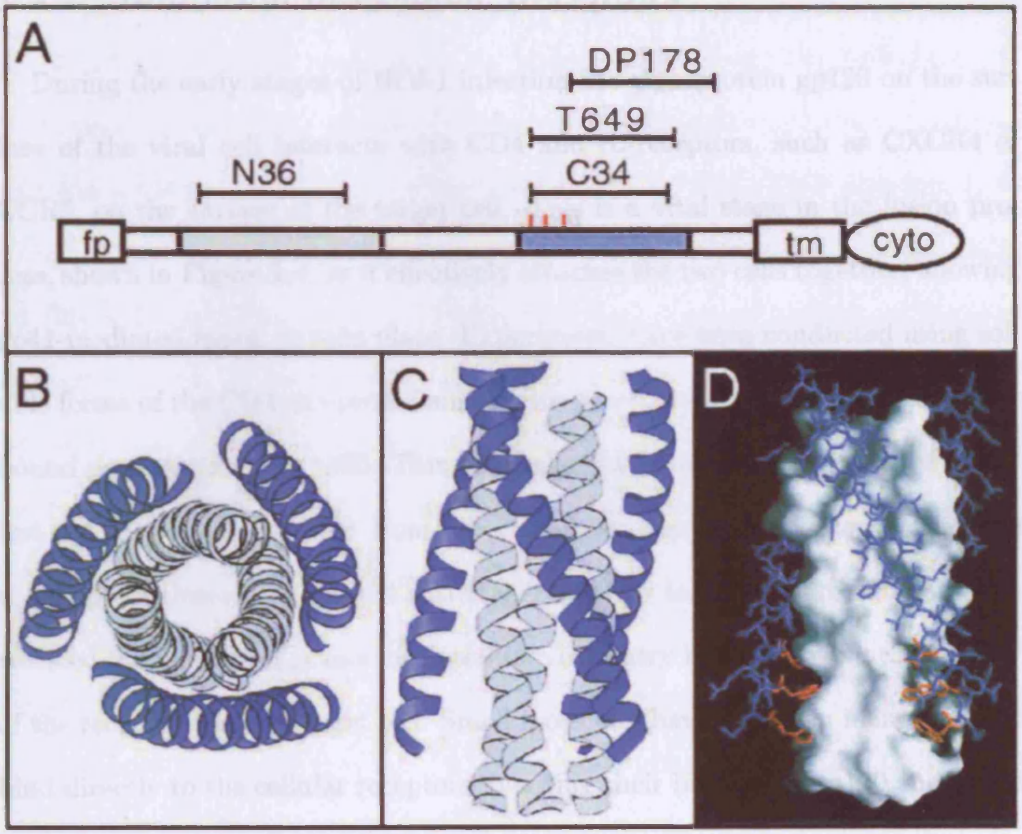

Figure 1.5: The structure of gp41. (A) Schematic showing the N-peptide and C-peptide regions, including the fusion peptide (fp), transmembrane segment (tm) and cytoplasmic region (cyto). The amino acid sequences representing some HIV-1 fusion inhibitors are also illustrated. The location of key pocket-binding residues Trp-628, Trp-631 and Ile-635 are indicated by red bars. (B,C) Helical representations of the packing of the N-peptides (grey) and C-peptides (blue) in the fusogenic conformation. (D) Illustration of the hydrophobic grooves and pockets on the surface of the N-peptide trimeric core. The pocket-binding residues of C34 are shown in red. Taken from Chan and Kim.[17]

1.4.2 Inhibition of Cellular Receptors

During the early stages of HIV-1 infection the glycoprotein gp120 on the surface of the viral cell interacts with CD4 and co-receptors, such as CXCR4 or CCR5, on the surface of the target cell. This is a vital stage in the fusion process, shown in Figure 1.4, as it effectively attaches the two cells together, allowing gp41-mediated fusion to take place. Experiments have been conducted using soluble forms of the CD4 receptors found on the target cell that suggested that they bound preferentially to gp120. This was successful for laboratory strains of HIV-1 but not for viral cells taken from other sources. More recently, variants of the co-receptors themselves showed antiviral activity by binding to gp120. They also reduced the number of points for potential viral entry by suppressing the activity of the receptors on the target cell. Small molecules have also been identified that bind directly to the cellular receptors, blocking their binding to gp120 and hence inhibiting viral entry.[34]

1.4.3 Viral Entry Inhibitors

1.4.3.1 C-Peptides

It has been shown that peptides derived from the C-terminal helix of gp41 can inhibit gp41-mediated fusion by binding to the transiently-exposed pre-hairpin intermediate, preventing the C-peptide region of the bound gp41 from interacting with the N-terminal core.[26–28,35–37] This type of binding is described as dominant-negative and prevents the essential conformational change to the fusogenic form of gp41.

Several C-peptides have been investigated for inhibitory activity, such as C34, T649 and DP-178 (T20) and these are illustrated schematically in Figure 1.5. Of these, C34 and T649 contain residues that bind to the hydrophobic groove and the pocket on the surface of the core, while DP-178 (T20) does not bind to the pocket but still exhibits inhibitory activity. However, the measured IC_{50} for C34 is 2–5 nM compared to 50–60 nM for T20, strongly suggesting that the interactions between the pocket-binding residues and the pockets on the surface of the N-peptide core play a vital role in determining the potency of the HIV-1 entry inhibitor.[20] Furthermore, since the residues forming the hydrophobic pockets are essentially conserved in the vast majority of strains of HIV-1, the virus is less likely to develop strains that are resistant to inhibitors that bind specifically to the pockets, rather than binding solely to the hydrophobic grooves.[21] It also suggests that the pocket itself could be a valid target for HIV entry inhibitors.[36]

Shorter C-peptides have also been investigated for inhibitory activity. C-peptides of 19 residues in length were found to have poor inhibitory activity[38] while C-peptides consisting of fourteen residues showed no inhibitory activity at μM concentrations.[39] In general, C-peptides are only α-helical when binding to the trimeric core but are found as random coils in solution.[37,40] Thus, a large entropic penalty is incurred upon complexation as the C-peptide can no longer occupy such a large number of possible non-helical conformations but is restricted to the α-helical structure.[40] This effect is particularly significant for short C-peptides as they only bind through a relatively small interface with the N-peptide core, resulting in a smaller change in enthalpy upon complexation.

The inhibitory activity of short C-peptides was investigated by using tech-

niques to restrain the C-peptides using cross-linking bridges between two residues and substituting for non-natural amino acids that have a higher propensity for forming regular helical structures in solution.[40] It was shown that by reducing the entropic cost of association, peptides such as C14 could be modified to inhibit cellular fusion.[40] The modified peptide that showed the best inhibitory activity was called C14linkmid (C14lm) and this peptide is used as the basis for the calculations on C-peptide inhibitors in this work. C14linkmid has a diaminoalkane bridge linking two amino acids separated by seven residues and a 12-residue version of the peptide, C12, is described in more detail in Chapter 4. When the linker was broken to form C14unlinkmid (C14unlm), the inhibitory activity was reduced significantly.[40]

1.4.3.2 D-Peptides

The importance of binding to the hydrophobic pocket for inhibitory activity makes it a target for small molecules that can disrupt the conformational change to the fusogenic form of gp41. Small, mirror-image D-peptides have been identified that bind exclusively to the pocket and inhibit gp41-mediated fusion.[21,41]

Eckert *et al.*[21] began by designing a peptide to represent the trimeric N-peptide core of gp41. This was achieved by fusing a soluble trimeric coiled-coil from a peptide called GCN4-pI_QI[42] to the portion of N36 that comprises the hydrophobic pocket. This species formed aggregates of order greater than the desired three but a small number of mutations were made to improve solubility and the resulting species formed fully α-helical trimers similar to the N-peptide trimeric core of gp41. This peptide was designated IQN17 and its mirror-image

D-IQN17 was used in a mirror-image phage display screening[43] to find L-peptides that bound exclusively to the pocket. The peptides that bound to D-IQN17 were also tested against control peptides that lacked the hydrophobic pocket residues, in some cases exhibiting one thousand times stronger binding to the peptide containing the pocket residues.[21]

Nine L-peptides were identified and their mirror-image D-peptides were synthesised and tested for inhibitory activity. Eight of the nine peptides shared a common sequence pattern that included a disulphide bridge between two cysteine residues, making the peptides roughly circular in shape. These eight D-peptides inhibited HIV cellular fusion and infection at μM concentrations, while the ninth D-peptide exhibited no inhibitory activity.[21] The amino acid sequences of the eight inhibitory D-peptides are given in Table 1.1.

These D-peptide inhibitors are 16 or 18 residues in length compared to 30 or more for the C-peptide inhibitors discussed previously. In order for drugs to be orally bioavailable they must be able to cross cell membranes readily and must therefore be below a certain size and although these D-peptides are still too large, they could act as a starting place in the search for smaller and more potent HIV-1 inhibitors. Only six of the residues make contact with the hydrophobic cavity and mutations could be possible to reduce the size of the peptides. As suggested previously, the conserved nature of the hydrophobic pocket residues in almost all of the reported strains of HIV-1 means that there would be a reduced risk of the virus mutating rapidly to become resistant to drugs based on these D-peptides.[21] Furthermore, due to their mirror-image nature, these D-peptides would be expected to be resistant to degradation by natural L-chirality enzymes

D-peptide	Amino Acid Sequence	Total Charge	Syncytia $IC_{50}/\mu M$	Entry $IC_{50}/\mu M$
	GACxxxxxEWxWLCAA			
D10-p1	GACEARHREWAWLCAA	0	29	
D10-p1-2K	KKGACEARHREWAWLCAA	+2	46	85
D10-p3	GACGLGQEEWFWLCAA	-2		
D10-p3-2K	KKGACGLGQEEWFWLCAA	0	16	39
D10-p4	GACDLKAKEWFWLCAA	0	64	
D10-p4-2K	KKGACDLKAKEWFWLCAA	+2	83	160
D10-p5	GACELLGWEWAWLCAA	-2		
D10-p5-2K	KKGACELLGWEWAWLCAA	0	3.6	11
D10-p6	GACSRSQPEWEWLCAA	-1	53	94
D10-p6-2K	KKGACSRSQPEWEWLCAA	+1	113	210
D10-p7	GACLLRAPEWGWLCAA	0		
D10-p7-2K	KKGACLLRAPEWGWLCAA	+2	41	
D10-p10	GACMRGEWEWSWLCAA	-1		
D10-p10-2K	KKGACMRGEWEWSWLCAA	+1	48	96
D10-p12	GACPPLNKEWAWLCAA	-2		
D10-p12-2K	KKGACPPLNKEWAWLCAA	0	130	270

Table 1.1: Amino acid sequences, total charges and IC_{50} values of inhibitory D-peptides for syncytia formation and viral entry.[21] The D-peptides were also synthesised with two additional lysine residues to improve solubility.

within the body.[21]

1.4.3.3 N-Peptides

In the same way that C-peptides can inhibit HIV fusion by binding to the pre-hairpin intermediate and preventing the conformational change to the fusogenic

form of gp41, peptides derived from the N-terminal region of the protein might be expected to inhibit fusion by binding to the exposed C-peptide. However, N-peptides in aqueous solution tend to form high-order aggregates, reducing the possibility of binding to the transiently-exposed intermediate.[19] The high-order aggregation of N-peptides has been investigated and shown to be the cause of the lack of inhibitory activity.[44] When the N-peptides are fused to other proteins that limit the aggregation to trimeric species, as was done in the design of IQN17, similar inhibitory activity to that exhibited by C-peptides is observed.

A small protein, designated 5-Helix, has also been designed,[19] consisting of three C-peptide (C38) helices and two N-peptide (N40) helices. These are joined together by linkers to form a single protein that resembles the six-helical bundle of the fusogenic conformation of gp41 but with one of the C-peptide helices removed, as shown in Figure 1.6. This provides a potential binding site for the exposed C-peptide of the pre-hairpin intermediate and 5-helix has been found to inhibit HIV fusion at similar concentrations to the C-peptides discussed previously. A similar protein, 6-Helix, was also engineered that did not contain the potential binding site for the C-peptide helix. This protein did not inhibit cellular fusion.[19] This is further evidence that the proposed mechanism for viral fusion is correct, again confirming the pre-hairpin intermediate as a potential target for anti-HIV drugs.

Peptides derived from the N-terminal region with point mutations have also been shown to inhibit gp41-mediated fusion. By making substitutions designed to block formation of the trimeric coiled-coil core while preserving the interactions with the C-terminal region, it was shown that the formation of the trimeric coiled-

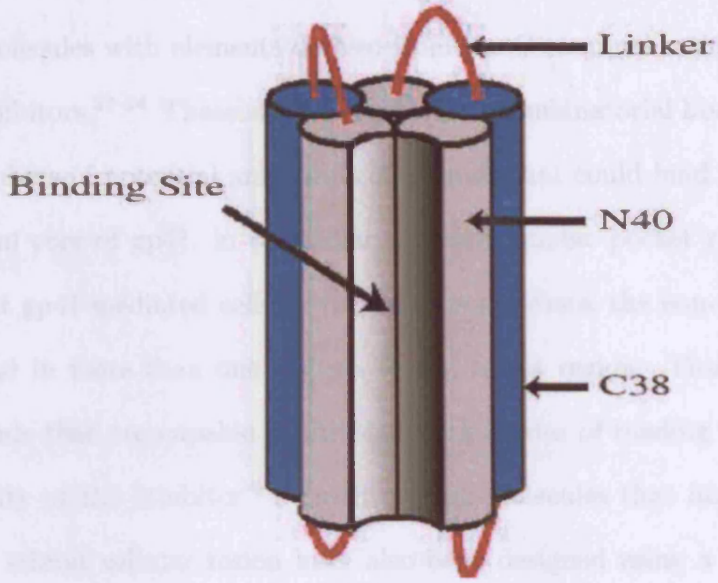

Figure 1.6: Schematic diagram of the structure of 5-Helix showing the C-peptide binding site and C-peptide–N-peptide linkers.[19]

coil was essential for binding to the C-terminal region. However, when mutations were made that allowed the formation of the core but decreased the interactions with the C-terminal region, inhibitory activity was still seen.[45] This was ascribed to the ability of the mutant N-peptide to disrupt the formation of the N-terminal trimeric core in the native gp41.[45]

1.4.3.4 Non-Peptide Inhibitors

Many of the HIV-1 entry inhibitors described previously are too large to be orally bioavailable. Furthermore, peptides are often susceptible to enzymatic degradation in the body, making them difficult to use as potential drugs. Several

studies have been carried out in order to find non-peptide ligands that could form hybrid molecules with elements derived from the C-terminal region of gp41 that act as inhibitors.[22–24] These studies used biased combinatorial libraries to screen large numbers of potential anti-viral compounds that could bind to the trimeric N-terminal core of gp41, in particular the hydrophobic pocket region, in order to prevent gp41-mediated cellular fusion. In some cases, the non-peptide moiety could bind in more than one fashion to the target region. This could lead to new ligands that are capable of utilising both modes of binding and improving the activity of the inhibitor.[24] Small organic molecules that bind to the gp41 core and inhibit cellular fusion have also been designed using a computational docking scheme. Some of these molecules were found to inhibit cellular fusion at μM concentrations.[46]

Chapter 2

Biomolecular Simulation

2.1 Introduction

As discussed in the previous chapter, a vast number of physiological processes are governed by interactions between biological molecules such as proteins and nucleic acids. Theoretical methods of studying the behaviour of such molecules have become increasingly common as computational resources have increased in recent times, allowing more and more complex biomolecular systems to be studied.[1,47–58] Different aspects of biological activity require the use of different theoretical methods, requiring that a range of techniques be used to probe the behaviour of these macromolecules. For example, quantum mechanical techniques are necessary when studying any process where chemical bonds are formed or broken[59–61] but classical techniques have been shown to perform well for processes involving non-covalent aggregation of proteins and nucleic acids.[1,62–66] These classical techniques and their application to free energy calculations are discussed

in some detail in this chapter.

One of the key assumptions used in almost all biomolecular simulations is the Born-Oppenheimer approximation, which states that the electrons and nuclei move in different time scales, allowing the nuclei to be treated as essentially fixed when considering the motion of the electrons.[67] The simulation can then be considered in two stages: determining the electronic potential energy surface and simulating the dynamical behaviour of the nuclei.

One of the most important objectives of quantum chemistry is to solve the non-relativistic, time-independent Schrödinger equation, in order to determine the electronic structure of a molecular system. The two main approaches to this are *ab initio* techniques, in which a model wavefunction is chosen and the Schrödinger equation is solved without recourse to empirical parameters, and semi-empirical techniques, which make use of simplified forms of the Hamiltonian and use some adjustable empirical parameters to reduce the computational cost of the calculation.[60,67,68]

Another technique that has been particularly successful for large, biologically significant molecules such as proteins and nucleic acids does not involve solving the Schrödinger equation at all, but uses a classical molecular mechanical force field to represent the potential energy surface. The potential energy has contributions due to bond stretching, bending and twisting, as well as van der Waals and electrostatic interactions between atoms. Other terms can be included, dependent on the individual force field and molecular system studied. Each force field has many empirical parameters that are obtained from comparison with experiments and high level *ab initio* calculations, meaning that care must be taken

when selecting a force field for a given molecular system.[1,63,64,68]

Each technique has advantages, such as the computational efficiency of the force field methods compared to *ab initio* and semi-empirical techniques, and disadvantages, including the simplifications and assumptions present in the force fields and their inability to treat bond formation or cleavage. In practice, many techniques have been developed that combine one or more of the above approaches to form a hybrid method, which treat the important part of a system, such as the active site of an enzyme, quantum mechanically and the rest of the system using a force field or similar technique.[69,70]

Two of the most widely used approaches to sampling the configurational phase space of large molecules are Monte Carlo (MC) and molecular dynamics (MD) techniques. MC techniques vary widely in their formulation and application but generally involve sampling a number of the possible accessible states of a system to obtain accurate statistical information.[68,71,72] MD techniques investigate the time-dependent behaviour of a system by solving Newton's equations of motion for the atoms, normally using a classical force field to determine the force on each atom.[1,63,64,68]

2.2 Representing the Potential Energy Surface

2.2.1 Molecular Mechanics Force Fields

Despite progress in electronic structure methods for biologically significant molecules, they are still very computationally expensive. It is common, therefore,

to use an entirely empirical molecular mechanical (MM) force field to model the potential energy surface of a molecular system.[1,63,64,68] One particular advantage of MM techniques is that they allow rapid evaluation of the potential energy function. However, they are incapable of simulating processes where chemical bonds are broken or formed, or where charge transfer takes place. Furthermore, great care must be taken in selecting a force field to use as each one is empirically parametrised for specific types of molecules.

In general, the potential energy is defined as having contributions from deformation of bond lengths, bond angles and torsional angles, as well as non-bonded interactions such as charge-charge and van der Waals interactions.[1,53] The functional forms of these energy terms are described in the Theory section of this work. This leads to a large number of empirical paramaters such as force constants, equilibrium bond lengths, equilibrium bond angles, equilibrium torsional angles, atomic partial charges and van der Waals radii. These parameters can be determined by experiment or by fitting to high-level *ab initio* calculations.[53]

Many force fields have been developed for simulation of proteins, including AMBER,[73,74] CHARMM,[75] CVFF,[76] GROMOS96,[77] OPLS-AA[78] and MM3PRO.[79] Some of these force fields also have implementations for the simulation of nucleic acids.[80] Force fields are also being developed that do not have an additive form for the potential energy, although the extra computational cost of these more accurate models could cause problems in biomolecular simulations.[53] MM force fields representing water, including TIP3P,[81] have been developed, allowing simulations in an explicitly modelled solvent environment. Methods combining MM with implicit solvent models are discussed later. The detailed

theory of MM methods are given in Chapter 3.

Evaluating the single-point potential energy of a molecular system, even a large, biologically significant one, is not very computationally expensive using empirical force fields. This means that large numbers of such evaluations can be performed in a short time, which is of particular importance when sampling large amounts of conformational phase space, for example during Monte Carlo (MC) or molecular dynamics (MD) simulations that can lead to an improved understanding of dynamical properties of biomolecules.[1,63,64,68]

2.2.2 Treatment of Solvation

Proteins and nucleic acids are not generally found in vacuum conditions and it is therefore necessary to consider the effect of the environment, usually aqueous solution, on the interactions within and between biomolecules. Several approaches to solvation are possible, ranging from treating both solute and solvent explicitly and atomistically to treating the solvent as an implicit continuous medium, with the same average properties as water, and even treating some of the solute degrees of freedom implicitly.[82]

The solvation free energy of a solute is usually divided into contributions due to electrostatic and non-polar solvation. The two most common methods used for evaluating the electrostatic or polar contribution are the Poisson-Boltzmann (PB)[82-84] and Generalised Born (GB)[85,86] methods, with the non-polar contribution usually treated as being proportional to the solvent accessible surface area (SASA).[82] These methods are discussed in more detail in Chapter 3. These methods have been used with MM force fields and MD simulations using explicit

water to evaluate free energies of binding and solvation, and these applications are discussed below in Section 2.4.2

The solvent can also be modelled explicitly using force fields such as TIP3P and TIP4P.[81] The parameters in these models are empirically determined to reproduce the enthalpy of vapourisation and density of water. The model TIP5P[87] has also been developed to reproduce the temperature dependence of the density of water. For most applications of explicit solvation to biomolecules, TIP3P is sufficient and is the force field used throughout this work.

2.3 Dynamical Methods

2.3.1 Monte Carlo Methods

Monte Carlo (MC) methods provide a way of sampling the conformational phase space of molecular systems. This is of particular importance in biologically significant systems as they are not stationary and the dynamical behaviour must be understood in order to gain insights into biological function.

An initial configuration is chosen for the system and its potential energy is evaluated, using one of the techniques described previously. A new configuration of the system is then generated, for example using the Metropolis method,[88] and the potential energy of the new state evaluated. If the new state has lower potential energy than the old one, i.e. $\Delta V < 0$ then the new state is accepted automatically. If the new state has higher potential energy than the initial state, i.e. $\Delta V > 0$, then it might still be accepted. A random number between zero

and unity is chosen and if this number is less than $\exp\left(\frac{-\Delta V}{kT}\right)$ then the new configuration is accepted and the process is repeated. Average values of molecular properties can then be generated from the sample configurations using as many steps as desired.[68]

Monte Carlo techniques have the advantage that only the potential energy for a configuration needs to be known. It is not necessary to evaluate gradients of the potential energy surface to generate the next configuration, as is the case for the molecular dynamics (MD) methods described below. It is also easy to constrain certain degrees of freedom when new configurations are generated. Structures that are far apart in configurational space can also be sampled rapidly using MC techniques, whereas MD methods tend to generate configurations that are close together, meaning that lengthy MD simulations are necessary for studying biomolecular processes. However, this also means that the methods for generating new configurations can be complicated, particularly for atomistic models of large molecules such as proteins and nucleic acids.[68]

MC techniques, in some cases combined with path integral (PI) theory, have been used to investigate the importance of quantum dynamical effects in proteins such as gelsolin, enkephalin and HIV-gp41.[89,90] These studies restricted the proteins to torsional motions and investigated the behaviour of the proteins at non-zero temperatures, finding that quantum dynamical effects could be significant for these types of molecules. A quantum thermal annealing technique using PIMC has been described for locating global minima of proteins.[91,92] Monte Carlo methods have also been applied to determining binding affinities for proteins, as discussed in further detail below.[53,93]

2.3.2 Molecular Dynamics Methods

Molecular dynamics (MD) methods also allow sampling of the conformational phase space of a system. The force on each particle is determined from the gradient of the potential energy surface, traditionally described using a molecular mechanical (MM) force field, and Newton's equations of motion are integrated numerically for each particle. Therefore, the time-dependent behaviour of a system can be simulated, with the only requirements being a starting structure and a suitable MM force field.[1,63,64,68] The details of the method are described in Chapter 3 and the simulation conditions used in this work are described in Chapter 4. The focus here is on the applications of MD techniques to systems of biological interest.

Early MD simulations of biomolecules were performed in vacuum,[94] providing insights into the importance of protein flexibility for biological function. However, it is now common to consider the influence of solvent, either explicitly or implicitly.[1,50] Counterions can also be added to the simulation, in order to make the system electronically neutral, and their effect on the trajectory considered.[95] MD simulations are computationally expensive, particularly when explicit solvent molecules are considered, meaning that a great deal of effort is put into developing new algorithms, for example to allow efficient treatment of electrostatic effects.[50,96–98]

Many biomolecules have been studied using MD, from the smallest alanine dipeptide in solution[99] to large systems such as the calcium-sensing protein calmodulin[100] in aqueous solution with sodium and chloride ions present. The calmodulin system had a total of over thirty thousand atoms and lasted for 3.0 ns,

requiring significant computational resources and illustrating the need for development of more efficient algorithms.[52] Calmodulin is a relatively small protein, consisting of 148 amino acid residues, and a large amount of the computational resources are spent determining detailed trajectories for the explicitly-modelled water molecules, which are not necessarily the most interesting parts of the simulation.

2.4 Statistical Mechanics

2.4.1 Ensemble Averages

MD simulations aim to give information on a macroscopic scale by simulating the microscopic behaviour of the atoms within molecules. Statistical techniques are used to provide the link between the microscopic and macroscopic levels.[101,102]

The thermodynamic or macroscopic state of a system is described in terms of a small number of parameters such as the temperature, pressure and number of atoms in the system. The mechanical or microscopic state of the system is defined in terms of the atomic positions and momenta, which can be considered as coordinates in $6N$-dimensional phase space. The state of the system is described by a single point in phase space and an ensemble is defined as a collection of points in phase space that all satisfy the same thermodynamic conditions. An MD simulation generates a series of points in phase space as a function of time. These points all belong to the same ensemble.[101]

In statistical mechanics, averages corresponding to experimentally observable

properties are defined in terms of ensemble averages. The ensemble average is represented by the notation $\langle\ \rangle$ and is given by:

$$\langle A\rangle_{\text{ensemble}} = \iint \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N A(\mathbf{p}^N, \mathbf{r}^N)\rho(\mathbf{p}^N, \mathbf{r}^N), \tag{2.1}$$

where A, the observable of interest, is a function of the momenta, $\mathbf{p}$, and the positions, $\mathbf{r}$, of the N atoms and ρ is the probability density of the ensemble, given by:

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \frac{1}{Z}\exp\left(\frac{-\mathscr{H}(\mathbf{p}^N, \mathbf{r}^N)}{RT}\right), \tag{2.2}$$

where $\mathscr{H}$ is the Hamiltonian, T is the temperature, R is the molar gas constant and Z is the total partition function, given by:

$$Z = \iint \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N \exp\left(\frac{-\mathscr{H}(\mathbf{p}^N, \mathbf{r}^N)}{RT}\right). \tag{2.3}$$

This integral is generally extremely difficult to calculate as all possible states of the system must be included. In order to calculate an ensemble average using MD, the simulation, which calculates points in the ensemble sequentially in time, must pass through all possible states corresponding to the thermodynamic conditions. It is easier to calculate a time average for the observable A, given by:

$$\langle A\rangle_{\text{time}} = \lim_{\tau\to\infty}\frac{1}{\tau}\int_{t=0}^{\tau} A(\mathbf{p}^N(t), \mathbf{r}^N(t))\mathrm{d}t \approx \frac{1}{M}\sum_{t=1}^{M} A(\mathbf{p}^N, \mathbf{r}^N), \tag{2.4}$$

where τ is the simulation time and M is the number of time steps.

If a system could be allowed to evolve indefinitely over time it would eventually sample all possible states. This leads to the so-called ergodic hypothesis, namely that the time average is the same as the ensemble average:[101]

$$\langle A\rangle_{\text{ensemble}} = \langle A\rangle_{\text{time}}. \tag{2.5}$$

This assumption is fundamental to the use of MD simulations to generate ensemble averages. It also follows that averages determined by MC simulations can be treated in the same way, provided that sufficient phase space sampling is ensured. For example, the average potential energy, V, for a system can be calculated as follows:

$$V = \langle V \rangle = \frac{1}{M} \sum_{i=1}^{M} V_i, \tag{2.6}$$

where M is the number of configurations or points sampled in phase space and V_i is the potential energy of configuration i. Another example is the average kinetic energy, K, given by:

$$K = \langle K \rangle = \frac{1}{M} \sum_{j=1}^{M} \left\{ \sum_{i=1}^{N} \frac{m_i}{2} v_i^2 \right\}_j, \tag{2.7}$$

where m_i is the mass of particle i and v_i is the velocity of particle i.

Clearly, it is vital for the MD simulation to be sufficiently long that enough possible configurations of the system are sampled in order that the ergodic hypothesis be valid. This represents one of the most significant problems in simulations of large biological systems.

2.4.2 Free Energy Calculations

2.4.2.1 Thermodynamic Cycle

Free energy is a state function, meaning that free energy differences can be calculated from a thermodynamic cycle similar to the one shown in Figure 2.1. The relative binding free energies, $\Delta\Delta G$, of two similar ligands, L1 and L2, bound

to the same protein, P, can be calculated:

$$\Delta\Delta G = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4, \tag{2.8}$$

where ΔG_1 and ΔG_2 are the binding free energies of L1 and L2, respectively, and ΔG_3 and ΔG_4 are the free energy changes upon transforming L1 to L2 in the free and bound states.[53,103,104] For the case where L1 and L2 are similar, it is normally easier to calculate these transformation free energies than the physical binding free energies.

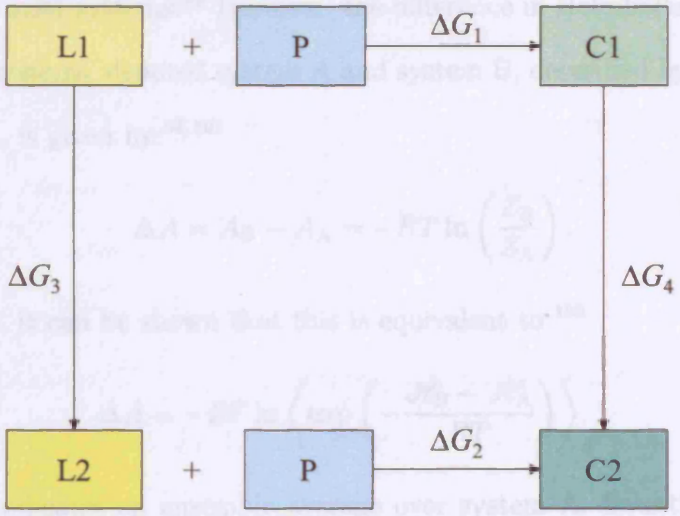

Figure 2.1: Thermodynamic cycle used for calculating differences in binding free energies for two similar ligands, L1 and L2, bound to the same protein, P.[53,103,104]

2.4.3 Free Energy Perturbation Methods

2.4.3.1 Theory

In the canonical ensemble, N, V and T are all constant, where N is the number of particles, V is the volume and T is the temperature. The Helmholtz free energy, A, of a system is given by:[101,102]

$$A = -RT \ln Z \tag{2.9}$$

where Z is the canonical partition function, as in equation 2.3. The evaluation of Z requires an integration over $3N$ degrees of freedom and is therefore not practical for most systems.[54] However, the difference in Helmholtz free energy of two related systems, denoted system A and system B, described by Hamiltonians $\mathscr{H}_A$ and $\mathscr{H}_B$, is given by:[54,103]

$$\Delta A = A_{\mathrm{B}} - A_{\mathrm{A}} = -RT \ln \left(\frac{Z_{\mathrm{B}}}{Z_{\mathrm{A}}} \right). \tag{2.10}$$

Furthermore, it can be shown that this is equivalent to:[103]

$$\Delta A = -RT \ln \left\langle \exp \left(-\frac{\mathscr{H}_{\mathrm{B}} - \mathscr{H}_{\mathrm{A}}}{RT} \right) \right\rangle_{\mathrm{A}}, \tag{2.11}$$

where $\langle\ \rangle_{\mathrm{A}}$ indicates an ensemble average over system A. Equation 2.11 is the basis of free energy perturbation (FEP) approaches but will only give a reasonable value for the free energy difference if the systems A and B differ only in a small way from one another. This problem is avoided by the introduction of a coupling parameter, λ, which can take values from zero to unity. The Hamiltonian, $\mathscr{H}(\lambda)$, is written as:[54]

$$\mathscr{H}(\lambda) = \lambda \mathscr{H}_{\mathrm{B}} + (1 - \lambda) \mathscr{H}_{\mathrm{A}}, \tag{2.12}$$

where $\lambda = 0$ gives $\mathscr{H}(\lambda) = \mathscr{H}_A$ and $\lambda = 1$ gives $\mathscr{H}(\lambda) = \mathscr{H}_B$. This allows the calculation to be broken into smaller parts as λ is varied from zero to unity. The interval in λ is chosen to be small enough that the free energy difference can be evaluated accurately at each stage. This is referred to as the free energy perturbation (FEP) method and equation 2.11 can be generalised to give:

$$\Delta A = -RT \sum_{i=1}^{N-1} \ln \left\langle \exp\left(-\frac{\mathscr{H}(\lambda_{i+1}) - \mathscr{H}(\lambda_i)}{RT} \right) \right\rangle_{\lambda_i}. \tag{2.13}$$

A similar method to FEP calculations is thermodynamic integration (TI), which involves calculating derivatives of the Hamiltonian at each value of λ and using numerical integration techniques to find the free energy difference:

$$\Delta A = \int_0^1 \left\langle \frac{\partial \mathscr{H}(\lambda)}{\partial \lambda} \right\rangle \mathrm{d}\lambda. \tag{2.14}$$

Both FEP and TI are rigorous methods that depend on the accuracy of the Hamiltonians used to represent the different systems. It is normal to neglect the kinetic energy contribution to the Hamiltonian and use $\mathscr{H} \approx V$, where V is the potential energy, usually represented by a MM force field. As mentioned previously, it is essential that enough conformational space is sampled that the ergodic hypothesis is valid. Therefore, these methods are computationally expensive.

The MM topologies can be handled in two distinct ways in free energy calculations.[54] In the first approach, the energy terms of the force field change the molecular topology as λ changes. In the second approach, two independent topologies are kept: for example, one topology for methanol and one for ethane in the calculation of the relative solvation free energy.[54,105] Both the methanol -OH group and the ethane -CH_3 group exist in the simulation but do not interact with one another. The interactions of these groups with their environments are

calculated using a variation of equation 2.12:

$$\mathscr{H}(\lambda) = \lambda^n \mathscr{H}_B + (1-\lambda)^n \mathscr{H}_A, \tag{2.15}$$

where the exponent n allows different pathways to be used in determining the free energies.[54] Within the single topology method, the λ-dependence is included more directly through the individual terms of the potential energy. It is also sometimes necessary to change the electrostatic part of the potential energy function on its own before changing the remainder of the Hamiltonian.[106,107] This is called electrostatic decoupling and is necessary when some atoms, particularly hydrogen, end up with small van der Waals repulsions but have some residual charge, giving an unrealistically high free energy change upon the approach of a solvent molecule.[54]

One potential disadvantage of FEP compared with TI is that the calculated free energy cannot be rigorously decomposed into its components,[108] since the logarithm of an exponential containing several terms is not the same as the sum of the logarithms of the individual components.[54] Once again, however, the main issues are the quality of the Hamiltonians and the degree of sampling rather than the choice of FEP or TI.

2.4.3.2 Applications

The concepts used in FEP calculations existed for some time before they were applied explicitly to chemical and biochemical systems.[109–111] Early applications to such systems included studies of solvation of noble gases[112] and solvation free energy contributions to a simple electron transfer reaction.[113] The potential of

FEP approaches was shown for model systems by Tembe and McCammon,[114] who first used the calculation of the non-physical legs of the thermodynamic cycle to evaluate $\Delta\Delta G$. Jorgensen and Ravimohan[105] used Monte Carlo (MC) and FEP techniques to calculate the relative solvation free energy of methanol and ethane, finding good agreement with experiments.

Early work on applying these methods to biologically significant systems by Bash *et al.*[106,107] studied the relative solvation free energy of a number of organic molecules, including amino acid side chains and nucleic acid bases, as well as the relative binding free energy of two ligands in complex with the enzyme thermolysin. Further work using FEP methods to investigate thermolysin inhibitors has been carried out by Merz and Kollman.[115] Warshel *et al.*[116] used MD and FEP to calculate the relative free energy of acid dissociation for key Asp and Glu residues in bovine pancreatic trypsin inhibitor (BPTI). An early application of FEP to protein-ligand association was carried out by Wong and McCammon,[117] who investigated the interactions between substituted benzamidines and trypsin. Simulations combining MC and FEP have also been utilised by Essex *et al.*[118] to predict the strongest of four trypsin inhibitors by studying trypsin–benzamidine complexes. These simulations also provided insights into the factors affecting the calculated relative binding free energies.

Free energy perturbation approaches have also been extensively applied to the inhibition of HIV protease (HIV-PR).[119–128] Reddy *et al.*[120] calculated the relative binding free energy of two peptide inhibitors, comprising seven and six amino acid residues, respectively. The large mutation from a seven-residue inhibitor to a six-residue inhibitor, involving the elimination of a valine residue, led to a re-

duction in the calculated binding free energy that was in excellent quantitative agreement with experiment. Rao *et al.*[123] studied two inhibitors of HIV-PR and found that constraining the active site of the enzyme to maintain its structure gave better agreement with experiments for the hydroxyethylene inhibitor (HEI). Another tight-binding hydroxyethylene inhibitor of HIV-PR, JG365, has been extensively investigated using FEP and MD, with several mutations being proposed that are predicted to improve the strength of the binding.[119,121,122] Cieplak[128] investigated the replacement of each peptide bond in JG365 with trans-ethylene or fluorethylene groups. Rick *et al.*[126] performed TI calculations to calculate relative binding free energies between a drug-resistant mutant of HIV-PR and the unmutated enzyme bound to several potent inhibitors. McCarrick and Kollman[127] used FEP calculations involving haloperidol thioketal (THK) and three of its derivatives bound to HIV-PR to predict further derivatives of THK with tighter binding.

Many other biological systems have been studied using FEP. Rastelli *et al.*[129] studied an inhibitor of aldose reductase and several analogous inhibitors involving point substitutions at the protein surface. The calculated relative binding free energies agreed qualitatively with experiments. Several MD/FEP studies of protein–carbohydrate complexes have been reported.[130–132] Zacharias *et al.*[132] compared the binding of arabinose and fucose to arabinose binding protein (ABP). Liang *et al.*[130] examined the relative binding of mannose and galactose to mannose binding protein (MBP). Pathiaseril and Woods[131] also found relative binding free energies in good agreement with available experimental data in their calculations on an anti-Salmonella antibody bound to variants of a trisaccharide

from Salmonella serotype B.

Fox *et al.*[133] used TI to calculate the relative binding free energy of two similar substrates with the antibody 17E8. Substitution of a -CH_2- group with a -S- reduced the calculated binding free energy by about 1 kcal mol^{-1}. Analysis of the components of the free energy indicated that the main factor in the preference for -CH_2- was the loss, upon complexation, of favourable solvation free energy for the more polar -S- side chain.

Simonson *et al.*[134] performed a detailed MD/TI study of the binding of aspartyl-tRNA synthetase (AspRS) to Asp and Asn residues. The specificity of binding to Asp over Asn is essential to maintaining the integrity of the genetic code. The calculated binding free energy was 15 kcal mol^{-1} more favourable for Asp binding to AspRS than Asn, largely due to the loss of electrostatic interactions with the carboxylate of the Asp side chain in the transformation to Asn. Lau and Karplus[135] had previously used TI in conjunction with MD to calculate the relative binding free energy of the Tyr substrate to wild type tyrosyl-tRNA synthetase (TyrRS) and a mutant of TyrRS, where a key Tyr residue had been replaced with Phe. The calculated free energy difference was attributed to interactions of the hydroxyl group of the wild type TyrRS and was found to be in good agreement with experimental values.

As well as calculating the relative binding free energy of similar ligands, FEP and TI have also been used to calculate absolute binding free energies for protein–ligand and DNA–ligand complexes. Helms and Wade[136] used MD with TI to calculate the absolute binding free energy of camphor with cytochrome P450cam from *Pseudonomas putida* to within 1 kcal mol^{-1} of the experimental value. This

was achieved by mutating the camphor, consisting of 11 non-hydrogen atoms, to six water molecules within the binding site.

Park and Lee[137] used MD and FEP to rationalise the relative binding strengths of several known histone deacetylase (HDAC) inhibitors. These inhibitors may provide a new approach to the treatment of cancer by preventing inappropriate uptake of HDACs by histone proteins complexed with DNA, which has been shown to affect gene transcription.[138] The results illustrated that any increase in favourable interactions between the enzyme and inhibitor, for example due to structural improvements in the inhibitor, must overcome the additional stability of the free modified inhibitor in solution to increase the binding free energy.

Saito and Sarai[139] calculated the effect on the binding free energy of DNA bound to λ-repressor of substituting thymine (T) with deoxyuracil (U) using FEP and MD simulations. The calculated relative binding free energy was in good agreement with available experimental data. Although T and U only differ in a small way, with a -CH group in T replaced by a hydrogen atom in U, the significant reduction in binding free energy was rationalised in terms of two effects: the lowering of the free energy of the unbound DNA as a result of a reduction in solvation free energy and the raising of the free energy of the bound state due to the loss of favourable van der Waals interactions in the binding site.

Setny and Geller[140] used MD simulations with the thermodynamic cycle approach to calculate binding free energy differences for adenosine triphosphate (ATP) and guanine triphosphate (GTP) to the protein casein kinase 2 (CK2). They obtained good agreement with experimental data and illustrated the importance of individual water molecules in the vicinity of the binding site.

2.4.4 Multiple Molecule Methods

2.4.4.1 Theory

FEP and TI are computationally expensive methods for calculating the free energy difference between two states or molecules. Methods such as the λ-dynamics of Kong and Brooks[141,142] and the adaptive chemical Monte Carlo/molecular dynamics (CMC/MD) method of Pitera and Kollman[143] have been developed in order to obtain the free energy differences between several similar states or molecules from a single simulation.

The λ-dynamics approach[141] treats λ as a set of variables, λ_j, where j can take values from 1 to n. Each of the n molecules is assigned a λ_j. As before, $\lambda_j = 0$ and $\lambda_j = 1$ correspond to initial and final states. An extended Hamiltonian is defined for the whole system, combining the Hamiltonians of the n molecules or states, a kinetic energy term for a set of fictitious masses and a potential function used for umbrella sampling. Umbrella sampling involves modifying the potential function to ensure that all states, even the most unfavourable ones, are sufficiently sampled during a simulation.[66,72] When the simulation reaches equilibrium, each of the n molecules has a different value of λ associated with it. The weighted histogram analysis method[144,145] (WHAM) can then be used to generate free energy profiles. λ-dynamics has the advantage over conventional FEP methods that rather than proceeding along a fixed path in λ-space, the approach dynamically searches for the optimal path in λ-space, allowing the simulations to converge more rapidly.[141]

The adaptive CMC/MD approach[143] uses the MD approach to sample the configurational space of each ligand and MC to sample the chemical space cor-

responding to the whole set of ligands being studied. The complex of one ligand and receptor provides the starting point for the MD simulation. After a certain amount of simulation time, a mutation to another of the set of ligands can occur, with the acceptance of any mutation determined by the usual MC criteria. At the end of the simulation, the ligands that bind most strongly to the receptor will have been sampled most often by the MC approach, allowing free energy differences between ligands to be obtained from an analysis of the sampling of each ligand.

2.4.4.2 Applications

The λ-dynamics approach was initially implemented in determining free energies of hydration for small organic molecules as a representative example to demonstrate the potential of the method.[141] Guo *et al.*[146] applied the method to a trypsin–inhibiitor system, obtaining results that were consistent with other free energy methods and experimental data. More recently, the method has been applied to determining the relative binding free energies of several inhibitors bound to the hepatitis C virus (HCV).[147] A single λ-dynamics simulation correctly identified strongly and weakly bound inhibitors. Multiple simulations were used and these gave accurate rankings for the potency of each of the inhibitors, with five of the seven inhibitors' calculated binding free energies found to be within 0.6 kcal mol^{-1} of the experimental value. Furthermore, the method was tested against conventional FEP calculations for six pairs of the ligands, with the free energy differences determined by λ-dynamics fitting better to experimental data for four of the six ligand pairs. The total cost of all the λ-dynamics simulations was

significantly less than the cost of the FEP calculations but obtained results that were comparable in accuracy.

Eriksson *et al.*[148] used the adaptive CMC/MD method to calculate the relative binding free energies for a series of HIV reverse transcriptase (HIV-RT) inhibitors. Again, the results were validated against conventional FEP/TI methods.

Pitera and Kollman[149] used the CMC/MD approach to calculate the relative free energies of several different amino acid side chains. The technique was used to calculate the relative solvation free energies of N-methylated and acetylated alanine, valine and serine. The calculations were in good agreement with TI calculations for these systems and illustrated the need to consider the flexibility of biological molecules and to include all of the signifcant conformations. They also used CMC/MD to investigate free energy changes due to several mutations at position 133 of T4 lysozyme from a single trajectory, again finding good agreement with previous free energy calculations[150] and experiment.

2.4.5 Pictorial Methods: PROFEC and OWFEG

2.4.5.1 Theory

Two methods have been proposed that use pictorial representations of the free energy surface around a potential inhibitor to suggest possible enhancements that would be expected to improve binding: pictorial representation of free energy components[151] (PROFEC) and the one window free energy grid[152] (OWFEG) method. Both methods involve mapping the system of interest onto a grid and approximately evaluating the free energy at each grid point.

In PROFEC, a neutral probe is placed at each grid point and a single window FEP calculation is performed. The free energies of the grid points are then shown as contours around the inhibitor, suggesting promising regions of the inhibitor for alterations to be made to improve the strength of binding.

OWFEG is similar to PROFEC but flexible regions of the ligand are explored by translating and rotating each grid point and its nearest atom. Furthermore, positively and negatively charged probes are used, as well as neutral ones, to determine the effect of adding charged groups at the various points of the grid.

2.4.5.2 Applications

Radmer and Kollman[151] used PROFEC in studying trypsin and its inhibition by benzamidine, finding it useful in ranking the strength of nine inhibitors. Lee and Kollman[153] combined PROFEC with FEP to predict more potent inhibitors of thymidylate synthase (TS), an enzyme whose inhibition could provide another way to treat cancer since it is involved in DNA synthesis. They predicted new, more potent TS inhibitors, modified from existing inhibitors, and validated their predictions using TI.

Pearlman[152] tested OWFEG on a protein–ligand complex, FKBP-12–FK506, and found excellent qualitative agreement with TI calculations. Pearlman and Charifson[154] introduced an OWFEG scoring grid, related to the free energy grid by a linear translation function with three fitted parameters. They then compared the ability of OWFEG with TI and several rapid scoring functions, designed for screening vast numbers of potential compounds, to predict the relative binding free energies of 16 inhibitors of a p38 kinase protein.[155] They defined a predictive

index (PI) that measures the reliability of a prediction method, with a PI of +1.0 being perfect correct prediction, 0.0 being random prediction and −1.0 being perfect incorrect prediction. The PI for TI was found to be 0.84 for the 16 p38 inhibitors, with OWFEG scoring 0.56. The rapid scoring functions had much lower PI values, even doing worse than random predictions in some cases. They concluded that precise free energy calculations such as FEP and TI allow accurate and reliable predictions that agree with experimental observations but much less computationally expensive methods such as OWFEG can give less reliable, but still qualitatively correct, predictions of relative binding free energies. Rapid scoring functions, designed as they are to be used in the rapid screening of large libraries of compounds, were not accurate enough to accurately predict the relative strengths of the inhibitors.

Eriksson *et al.*[148] used PROFEC in conjunction with the adaptive CMC/MD method and conventional FEP/TI methods. They used the technique to predict a new derivative of an inhibitor of HIV-RT that should bind more strongly and confirmed that its calculated binding free energy was 1–2 kcal mol^{-1} more favourable than the original inhibitor.

2.4.6 Linear Interaction Energy Method

2.4.6.1 Theory

The linear interaction energy (LIE) method for estimating absolute binding free energies was proposed by Åqvist *et al.*[156] The LIE method is based on linear response assumptions: in other words, the change in the polarisation of the

solvent caused by changes in the electrostatic field due to the solute is linear and characterised by a single dielectric constant. The absolute binding free energy, $\Delta G_{\text{binding}}$, for a protein–ligand complex is written as:

$$\Delta G_{\text{binding}} = \Delta G_{\text{binding}}^{\text{electrostatic}} + \Delta G_{\text{binding}}^{\text{vdW}}, \tag{2.16}$$

where $\Delta G_{\text{binding}}^{\text{electrostatic}}$ and $\Delta G_{\text{binding}}^{\text{vdW}}$ are the contributions due to electrostatic and van der Waals interactions, respectively. These components are estimated from ensemble averages, again denoted by $\langle\ \rangle$, taken from MD or MC simulations, as:

$$\Delta G_{\text{binding}} \approx \alpha \left\langle E_{\text{bound}}^{\text{electrostatic}} - E_{\text{free}}^{\text{electrostatic}} \right\rangle + \beta \left\langle E_{\text{bound}}^{\text{vdW}} - E_{\text{free}}^{\text{vdW}} \right\rangle, \tag{2.17}$$

where $E_{\text{bound}}^{\text{electrostatic}}$ and $E_{\text{bound}}^{\text{vdW}}$ are the electrostatic and van der Waals interaction energies, respectively, taken from a simulation of the bound protein–ligand complex; $E_{\text{free}}^{\text{electrostatic}}$ and $E_{\text{free}}^{\text{vdW}}$ are the electrostatic and van der Waals interaction energies, respectively, between the ligand and water, taken from simulations of the free ligand in water; and α and β are empirical parameters. A third parameter, γ, sometimes presented as being dependent on the solvent accessible surface area (SASA) change upon complexation, $\Delta\langle\text{SASA}\rangle$, has also been used to improve the fitting of binding free energies to experiment.[157–161]

2.4.6.2 Applications

Åqvist[156] applied the LIE method to calculating absolute and relative binding free energies of several inhibitors bound to endothiapepsin. For a calibration set of four inhibitors, $\alpha = 0.50$ and $\beta = 0.16$ gave a mean unsigned error of 0.4 kcal mol^{-1} for absolute binding free energies and 0.6 kcal mol^{-1} for relative binding free energies. These parameters were used to predict the binding free energy of

a fifth endothiapepsin inhibitor. The predicted value of -9.70 kcal mol^{-1} was found to be in good agreement with the experimental value of -9.84 kcal mol^{-1}. This method, using the same values of α and β and a correction term for the effects of long-range electrostatic contributions to the binding free energy, was also applied to HIV protease (HIV-PR) inhibitors[162] and benzamidine inhibitors of trypsin,[163] again finding good agreement with experimental studies. Hansson *et al.*[164] further refined the parameters using FEP calculations to determine the electrostatic coefficient, α. They found that a model where α took one of four values from 0.33 to 0.50, depending on the charged or dipolar groups in the ligand, $\beta = 0.18$ and $\gamma = 0$ gave optimal fitting to experiment.

Paulsen and Ornstein[165] found that a different value of the van der Waals parameter, $\beta = 1.043$, gave good agreement with experiment for 11 ligands bound to cytochrome P450cam. Wang *et al.*[166] found that $\beta = 1.0$ gave good agreement with experiment for 14 analogues of biotin bound to avidin, prompting Wang *et al.*[167] to investigate what factors determine the value of β in the LIE method. They examined several different protein–ligand complexes using MD simulations and defined a weighted non-polar desolvation ratio (WNDR), related to the hydrophobicity of the ligand and the binding site, that exhibited a linear correlation with β. The greater the proportion of non-polar groups that are buried in the binding site, the smaller the WNDR and hence the larger the value of β. Conversely, burying polar groups in the binding interface increases the WNDR, thereby reducing β. They found that using the WNDR, easily calculated for any protein–ligand system, generally predicted values of β that gave better estimates of the binding free energy for twelve neutral ligands bound to avidin than using

a fixed value, $\beta = 0.87$, taken from the best estimate of the binding free energy for the biotin–avidin system.

Jorgensen's group used the LIE method with MC to calculate hydration and binding free energies, adding a term, proportional to the change in solvent accessible surface area (SASA) upon complexation, to equation 2.17. Once again, the coefficients α and β were calibrated in a test set and used to calculate binding free energies of inhibitors of human thrombin[168] and FKBP12 inhibitors.[158]

Wall *et al.*[157] applied the LIE method with MC simulations to the binding of 15 inhibitors to the enzyme neuraminidase. They found that application of equation 2.17 led to root mean square (RMS) errors of 1.5 kcal mol^{-1}, compared to experiment, in the binding free energies. A detailed statistical analysis showed that the addition of the surface area-dependent parameter γ significantly improved the fit to experimental data. Ljungberg *et al.*[159] found that a constant γ was also necessary in their studies of the absolute binding energies of inhibitors binding to thrombin.

Almlöf *et al.*[161] recently investigated the dependence of the parameters α, β and γ in the LIE method on the force field used. They used three different force fields to investigate the binding of nine ligands to P450cam and found that the same parameters given by Hansson *et al.*[164] produced relative binding free energies in good agreement with experiment. However, for the absolute binding free energies, all three force fields predict values that are too positive. A value of $\gamma = -4.4$ kcal mol^{-1} brought the absolute binding free energies predicted by each of the three force fields to within 0.4 kcal mol^{-1}, suggesting that the parameters were not dependent on the choice of force field.

Chen *et al.*[169] used MD simulations with the LIE approach to calculate binding free energires for steroids binding to antibodies and predicted the binding free energy for an untested steroid. The LIE method was also used by van Lipzig *et al.*[170] to predict the binding affinites of ligands to the estrogen receptor (ER), also taking into account every possible orientation of the ligands within the binding site. Ersmark *et al.*[171] applied the method to the haemoglobin-degrading aspartic proteases, plasmepsin I and plasmepsin II, of the malaria parasite, to a series of potential inhibitors, reproducing the experimental binding data well. Asi *et al.*[172] applied the method using MD simulations to calculate binding free energies for wild type and mutant forms of the C-terminal domain of *Escherichia coli* (*E. coli.*) arginine repressor (ArgRc) bound to L-arginine and L-citrulline. Svab *et al.*[173] calculated binding free energies for sulfonamide inhibitors bound to matrix metalloproteinases, predicting which of two possible binding modes was favoured for the inhibitors.

2.4.7 MM–PBSA and MM–GBSA Methods

2.4.7.1 Theory

The theory of combining molecular mechanics (MM) and continuum solvation techniques, such as Poisson Boltzmann Surface Area (PBSA) and Generalised Born Surface Area (GBSA), is described in the Theory chapter of this thesis as it is directly used in the research reported here.

2.4.7.2 Applications

The approach of combining a force field with a continuum solvation model was introduced by Srinivasan *et al.*[174] to study the relative free energies of alternative forms of DNA in aqeuous solution. The MM–PBSA method has been used extensively due to its much lower computational cost compared with free energy perturbation methods. In particular, the work of Massova and Kollman[3] on the binding of the oncoprotein Mdm2 with the tumour suppressor peptide p53, in which they introduced the computational alanine scanning technique used in this work, has been discussed previously in Chapter 1.

Huo *et al.*[175] extended the method to investigate the binding of human growth hormone (HGH) and its receptoor, for which experimental alanine scanning data were available. The calculations were in good agreement with experiment for the 12 mutations, with the exception of two arginine mutations, for which a significant conformational change would take place upon replacement with alanine.

Chong *et al.*[176] illustrated the usefulness of the method for calculating relative binding free energies for germ line and mature forms of an antibody fragment to an anionic hapten. Reasonable agreement was also found for the absolute binding free energies. The stronger binding of the mature antibody was attributed to the optimisation of the binding site geometry and the resulting gain in favourable electrostatic interactions overcoming the loss of solvation free energy upon complexation.

Kuhn and Kollman[177] used MM–PBSA to study derivatives of biotin binding to avidin and streptavidin, reproducing relative binding free energies of 9-methylbiotin compounds. They used computational fluorine scanning, a tech-

nique similar to the computational alanine scanning described earlier, to replace hydrogen atoms in the MD trajectory with fluorine. This is a much cheaper alternative to traditional free energy perturbation methods as several potential mutations can be investigated from the same trajectory. Of the nine fluoro-substitutions investigated, only one was predicted to increase the binding free energy. The relative binding free energies were also in agreement with TI calculations.[177]

Kuhn and Kollman[178] also compared the MM–PBSA and LIE methods for calculating binding free energies for a number of diverse avidin and streptavidin ligands. They found that MM–PBSA performed better than LIE for a fixed value of $\beta = 1$. Although the LIE method allows for the empirical adjustment of β, the MM–PBSA approach introduces no empirical parameters that are not intrinsic to the force field.[53]

Wang and Kollman[179] studied the stability of HIV-PR dimers. Different dimers were ranked in order of binding free energy in good agreement with experiment. They introduced the method of virtual mutagenesis to suggest cavities near the binding interface of the dimer that could potentially be filled by a larger side chain in a mutated PR, which could then bind in a dominant-negative manner and inhibit dimer formation.

Santa *et al.*[180] have used MM–PBSA to investigate different conformations of tetrapeptides in order to gain insights into which types of turns are preferred in aqueous solution. Nordman *et al.*[181] have studied the binding of testosterone and several of its analogues to an anti-testosterone antibody. The method produced relative binding free energies in reasonable agreement with experiment for both

neutral and charged steroids. Computational mutagenesis techniques were used to rationalise the lower binding affinities of some ligands than testosterone in terms of the contributions from specific residues.

Several applications of the generalised Born surface area (GBSA) method, which is the method used in this work, have also been reported. Rizzo *et al.*[182] used the GBSA method, rather than explicit water molecules, for the MD simulation of six thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A, two homologous matrix metalloproteases (MMPs) that have been implicated in breast cancer. They also used MM–GBSA in post-proessing the trajectories and found a strong correlation between the calculated binding free energies and experiment, correctly predicting the preference in binding stromelysin-1 over gelatinase-A and rationalising this selectivity in terms of stronger van der Waals interactions and a smaller net desolvation penalty upon complexation.

Lepsik *et al.*[183] have also used MM–GBSA to study an inhibitor of wild-type and drug-resistant HIV-PR, designated QF34. Its binding to HIV-PR was compared with two clinically used drugs, saquinavir (SQV) and indinavir (IDV), and it was shown that its greater flexibility allowed QF34 to form more stable hydrogen bonds than either SQV or IDV. The MM–GBSA analysis showed that the van der Waals and electrostatic interactions favoured binding and that more residues of HIV-PR were involved in binding to the inhibitor than was the case for SQV or IDV.

Rinaldo *et al.*[184] have investigated several potential strontium-binding proteins, of interest in the area of nuclear waste management, using MM–GBSA. They analysed the dynamical and binding properties of seven mutants of a se-

quence from the calcium-binding protein calmodulin and identified factors that could contribute to a stronger specificty for strontium.

2.5 This Research

The aim of the research presented in this thesis is to investigate the non-covalent interactions between Mdm2 and p53, with particular emphasis on the contribution of each amino acid residue of the peptide to the overall binding free energy. The same approach is applied to the interactions of IQN17 with D10-p1, and IQN17 with two C-peptides, C12lm and C12unlm. Identification of the key peptide residues for binding to the protein could potentially lead to suggestions of changes in the HIV inhibitors that would improve the binding and hence reduce the concentration needed to successfully inhibit cellular fusion.

In this work, protein-peptide interactions are probed using MD simulations with explicitly modelled water. The AMBER[74] force field is used for the protein and peptide and the water molecules are represented using the TIP3P[81] model. The GBSA implicit solvation method is used in post-processing the MD trajectory to obtain solvation free energies for the complex, protein and peptide for each of the molecular systems studied. Each residue of the peptide is systematically mutated to alanine using a computer program developed in this laboratory and the effect on the binding free energy is evaluated. This allows the identification of the key residues in the binding between the protein and peptide. It also gives insights into the factors influencing binding, including whether electrostatic or van der Waals interactions are the main contribution to the binding energy from

a given peptide residue, or whether solvation effects dominate. The theory and methods used in this work are described in more detail in Chapters 3 and 4, respectively.

Part II

Theory and Methods

Chapter 3

Theory

3.1 Molecular Mechanics

3.1.1 Introduction

The time-independent behaviour of a molecule is described by the Schrödinger equation:[185,186]

$$\mathscr{H}\Psi = E\Psi, \tag{3.1}$$

where Ψ is the wavefunction of the system, E is the energy of the state and $\mathscr{H}$ is the Hamiltonian operator. The electrons in the system are much smaller and lighter than the nuclei. Thus, we can invoke the Born-Oppenheimer approximation and treat the motions of the electrons and the nuclei as separate. On the timescale of electronic motion, the nuclei can be treated as fixed, allowing the nuclear kinetic energy and internuclear repulsion terms to be neglected. Thus,

the Schrödinger equation can be expressed as:

$$\mathscr{H}_{electronic}\psi = E_{electronic}\psi, \tag{3.2}$$

where ψ is the electronic wavefunction and $\mathscr{H}_{electronic}$, the electronic Hamiltonian operator, is given by:

$$\mathscr{H}_{electronic} = \sum_{i=1}^{n} \frac{-\hbar^2}{2m}\nabla_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{N} \frac{z_i e^2}{4\pi\epsilon_0 \mid \mathbf{r}_i - \mathbf{R}_j \mid} + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{e^2}{4\pi\epsilon_0 \mid \mathbf{r}_i - \mathbf{r}_j \mid}, \tag{3.3}$$

where ∇ is the differential operator, N is the number of nuclei, n is the number of electrons, m_e is the mass of an electron, q_i is the charge on particle i, ϵ_0 is the permittivity of free space, $\hbar$ is Planck's constant divided by 2π, $\mathbf{r}_i$ is the position of electron i and $\mathbf{R}_j$ is the position of nucleus j.

These three terms in the Hamiltonian represent the kinetic energy of the electrons, the attraction between the electrons and the fixed nuclei and the repulsion between the electrons, respectively. Despite progress in electronic structure theory, it is still not possible to solve the Schrödinger equation accurately and efficiently for large molecules. Therefore, it is often necessary to use simplified models such as molecular mechanics (MM) to study the interactions within proteins and polypeptides.

Molecular Mechanics[63–65,68] (MM) is an atomistic model that considers the interactions between individual atoms using classical mechanics. These interactions can be broadly divided into bonded or internal interactions and non-bonded interactions. Many molecular mechanical force fields have been developed such as CHARMM,[75] MM2,[187] MM3,[188] MM3PRO[79] and OPLSAA.[78] The AMBER[73,189] force field has been used throughout this work and its functional form

is described below.

3.1.2 Internal Interactions

There are three main types of internal motion that must be considered: bonds stretching and compressing, bond angles deforming and torsional motions. The potential energy due to deformation of bond lengths, E_{bonds}, and bond angles, E_{angles}, away from their equilibrium values are modelled within the AMBER[73,189] force field by harmonic terms as shown in equations 3.4 and 3.5:

$$E_{\text{bonds}} = \sum_{\text{bonds}} \frac{1}{2} K_r \left(r - r_{\text{eq}}\right)^2, \tag{3.4}$$

$$E_{\text{angles}} = \sum_{\text{angles}} \frac{1}{2} K_\theta \left(\theta - \theta_{\text{eq}}\right)^2, \tag{3.5}$$

where K_r and K_θ are force constants, r and θ are bond lengths and angles, respectively, and r_{eq} and θ_{eq} are equilibrium values of the bond lengths and angles, respectively. K_r, K_θ, r_{eq} and θ_{eq} are parameters that are established within the force field by fitting to experimental data.

The potential energy due to torsional motions, E_{torsions}, is given by:

$$E_{\text{torsions}} = \sum_{\text{torsions}} \frac{1}{2} V_n \left[1 + \cos(n\phi - \gamma)\right], \tag{3.6}$$

where ϕ are the dihedral angles, V_n is a force constant and γ is a phase term.

The sum of the energies due to these internal motions is the internal energy, E_{internal}, given by:

$$E_{\text{internal}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}}. \tag{3.7}$$

3.1.3 Non-bonded Interactions

The AMBER[74] force field assigns partial charges to each type of atom and the energy due to electrostatic interactions between atoms, $E_{\text{electrostatic}}$, is given by a Coulombic term:

$$E_{\text{electrostatic}} = \sum_{i<j} \frac{q_i q_j}{\epsilon r_{ij}}, \tag{3.8}$$

where q_i is the charge on atom i, ϵ is the permittivity or dielectric constant and r_{ij} is the distance between atoms i and j.

The energy due to van der Waals interactions, E_{vdW}, is modelled by a Lennard-Jones potential:

$$E_{\text{vdW}} = \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), \tag{3.9}$$

where A_{ij} and B_{ij} are parameters that characterise the interaction between two given types of atom.

Energies due to non-bonded interactions are only evaluated for atoms separated by at least three bonds. Interactions between atoms separated by exactly three bonds, i.e. 1–4 interactions, are evaluated but a scaling factor is applied to the energy terms. For electrostatic interactions, the calculated energy is divided by a factor of 2.0. For van der Waals interactions the scaling factor is 1.2.

3.1.4 Gas Phase Energy

The gas phase energy, E_{gas}, of a molecule is defined in terms of internal and non-bonded energies as shown in equation 3.10.

$$E_{\text{gas}} = E_{\text{internal}} + E_{\text{electrostatic}} + E_{\text{vdW}} \tag{3.10}$$

3.2 Implicit Solvent Treatment

3.2.1 Introduction

Proteins are usually found in aqueous solution and it is therefore essential to consider the effect of solvent when studying protein interactions. It is possible to model the solvent atoms explicitly using a potential energy surface such as TIP3P,[81] as has been done during the molecular dynamics simulations in this work. However, this is computationally expensive as large amounts of time are spent calculating the behaviour of individual solvent atoms when it is actually the effect of these atoms on the solute that are of primary interest. Furthermore, simulations with explicit solvent molecules have inherent assumptions involving the treatment of long range electrostatic interactions.[82] Therefore, it is becoming increasingly common to treat the solvent as an implicit continuous medium with the same average properties as the real solvent.[3,56,82,85,86,99,174,175,190–199] The free energy of solvation of a molecular species in this solvent continuum is defined in terms of a solvent-solvent cavity term, G_{cav}, a solute-solvent van der Waals term, G_{vdW}, and a solute-solvent electrostatic polarisation term, G_{polar}, as follows:

$$G_{\mathrm{solvation}} = G_{\mathrm{cav}} + G_{\mathrm{vdW}} + G_{\mathrm{polar}}. \tag{3.11}$$

Solvating a molecule in such a continuous solvent medium can be thought of as a three-stage process. First, a cavity is made in the solvent and the solute is inserted but does not interact with the solvent. Next, van der Waals interactions between the solute and solvent are switched on and, finally, the solute and solvent are allowed to interact electrostatically.[68,82,85,86,198] Normally, the

first two stages of this process are considered simultaneously and the solute is essentially inserted into the solvent with no partial charges and then the charges are switched on. Hence, it is normal to define $G_{\text{solvation}}$ in terms of polar and non-polar contributions:

$$G_{\text{solvation}} = G_{\text{non-polar}} + G_{\text{polar}}. \tag{3.12}$$

3.2.2 Surface Area Methods

The free energy of aqueous solvation for non-polar solutes such as saturated hydrocarbons is approximately proportional to their solvent accessible surface area (SASA).[200–202] Hence, it is normal to estimate the non-polar contribution to the solvation free energy[203,204] using an expression such as:

$$G_{\text{non-polar}} = \sum_k \sigma_k \text{SASA}_k, \tag{3.13}$$

where SASA_k is the total solvent accessible surface area of all atoms of type k, which can be calculated numerically[205,206] and σ_k is an empirically determined atomic solvation parameter. Clearly, for saturated hydrocarbons, since $G_{\text{polar}} \sim 0$, this is sufficient to obtain good estimates for the free energy of solvation. However, for molecules such as proteins, it is also necessary to determine the polar contributions to the solvation energy.

3.2.3 Poisson-Boltzmann Method

The classical treatment of electrostatic interactions in solution is based on the Poisson-Boltzmann equation:[56,82,83]

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \Phi(\mathbf{r})] - \kappa^2(\mathbf{r}) \sinh(\Phi(\mathbf{r})) + 4\pi\rho(\mathbf{r}) = 0, \tag{3.14}$$

where $\Phi(\mathbf{r})$ is the electrostatic potential, $\epsilon(\mathbf{r})$ is the dielectric constant, $\rho(\mathbf{r})$ is the charge density, $\kappa(\mathbf{r})$ is the Debye-Hückel parameter and ∇ is the differential operator. Note that all of these terms depend on position. The second term in equation 3.14 accounts for salt effects due to mobile ions in solution. If $\Phi(\mathbf{r})$ is small, the sinh term can be expanded as a Taylor series[207] and truncated at the first term to yield the linear Poisson-Boltzmann equation. Furthermore, when such salt effects are absent, i.e. when $\kappa(\mathbf{r}) = 0$, equation 3.14 reduces to the Poisson equation:[68,83,84]

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}). \tag{3.15}$$

It should also be noted that this in turn reduces to Coulomb's law if the dielectric constant is the same throughout all space.[83] The charge density, $\rho(\mathbf{r})$, is determined by the spatial distribution of partial atomic charges within the solute. The most difficult term to describe accurately is the dielectric constant, $\epsilon(\mathbf{r})$. Due to the inhomogeneous nature of proteins, there is no one such constant that can describe the dielectric response of an entire protein to an applied electric field. For example, in the protein interior where hydrophobic residues are most common and the response to any electric field is small, a dielectric constant of 2–4 is appropriate. However, at the protein surface where polar or charged residues can produce a significant dielectric response to the presence of an applied electric field, a value for the dieletric constant of 10–30 is more appropriate.[56] Hence, it is necessary to consider the dielectric constant as an empirical parameter that attempts to describe the overall dielectric behaviour.[56]

The Poisson-Boltzmann equation is generally solved numerically due to the

complicated shapes of protein surfaces. The most common methods involve placing the protein at the centre of a cubic grid and solving the equation iteratively until a preset level of tolerance is met at all grid points. This can be done using programs such as DelPhi,[208] which can be used with common molecular mechanical force fields. Solving the Poisson-Boltzmann equation yields the electrostatic potential, $\Phi(\mathbf{r})$, throughout the space in and around the macromolecule of interest. For the case of the linear Poisson-Boltzmann equation, the total electrostatic free energy of the system, G_{es}, is simply given by:

$$G_{es} = \frac{1}{2}\sum_i q_i \Phi_i, \tag{3.16}$$

where Φ_i is the electrostatic potential at the position of charge i.[56]

The polar contribution to the solvation free energy, G_{polar}, is thus the difference between the electrostatic free energy in solution, given by equation 3.16, and the electrostatic free energy in vacuo, given by Coulomb's law, as shown in equation 3.8.

Solving the Poisson-Boltzmann equation in this fashion is computationally expensive and the method is therefore best used for evaluating G_{polar} for single structures and a faster method for evaluating G_{polar} is required for calculations involving large numbers of structures. Simple models involving distance-dependent dielectric constants, $\epsilon(r) \propto r$, have been employed but are of dubious accuracy and more precise models based on the Born expression are commonly used.

3.2.4 Generalised Born Method

The Born expression[209] for a spherical ion is given by:

$$G_{\text{Born}} \propto \left(1 - \frac{1}{\epsilon}\right) \frac{q^2}{2\alpha}, \tag{3.17}$$

where q is the charge on the ion, α is the effective Born radius and ϵ is the dielectric constant of the medium.

The total electrostatic free energy of a system of n widely separated particles with separations r, in solution, G_{es}, can be written as:

$$G_{\text{es}} \propto \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}\epsilon} - \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \frac{q_i^2}{\alpha_i}, \tag{3.18}$$

where the first term in this equation is Coulomb's law in a dielectric medium. This can be expanded into two terms: Coulomb's law in a vacuum and a term that resembles the Born expression. This gives the following expression:

$$G_{\text{es}} \propto \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}} - \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}} - \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \frac{q_i^2}{\alpha_i}. \tag{3.19}$$

Therefore, G_{polar} can be written in terms of some function f_{GB} in the Generalised Born (GB) equation:

$$G_{\text{polar}} \propto \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{q_i q_j}{f_{\text{GB}}}, \tag{3.20}$$

where f_{GB} is a function of r_{ij} and α_i and is not defined uniquely but is chosen to make equation 3.20 mimic the equations of classical electrostatics.[86] One effective expression for f_{GB} is:

$$f_{\text{GB}} = \sqrt{\left(r_{ij}^2 + \alpha_{ij}^2 \exp\left(-\frac{r_{ij}^2}{(2\alpha_{ij})^2}\right)\right)},$$

where $\alpha_{ij} = \sqrt{\alpha_i \alpha_j}$. As r_{ij} becomes larger than α_{ij} the exponential term becomes very small. In other words, the exponential term forces the expression for G_{polar} to approximate the dielectric part of Coulomb's law as the two atoms i and j move beyond the contact distance of their Born radii.[86]

The Generalised Born equation, equation 3.20, requires a Born radius α_i for each atom, which depends on the positions and sizes of all the other atoms in the solute as they displace the dielectric medium. The Born radius for an atom is more like an average distance from the partial charge to the boundary of the dielectric medium.[86] These atomic Born radii can be evaluated accurately using numerical techniques[85] but this is time consuming. A fast analytical approach to calculating Born radii has been introduced by Qiu *et al.*[86] and this method is used throughout this work for calculating polar contributions to the solvation free energy.

The general method for computing Born radii starts with the Born expression, equation 3.17, and considers a polyatomic solute with no partial charges. For this system G_{polar} is zero but the dielectric medium is displaced by the solute atoms. A charge q_i can then be introduced on atom i and the Born equation can be used to calculate an average effective Born radius, α_i, provided that G_{polar} can be evaluated for the hypothetical system. Qiu *et al.*[86] developed an analytical function for calculating G_{polar} that estimated the effect of the presence of each of the other atoms in the solute and their displacement of the dielectric medium. Their effect on G_{polar} is based on a V/r_{ij}^4 relationship between the volume of dielectric medium displaced by atom j and the separation of atoms i and j. This represents the loss of the interaction between the atomic partial charge, q_i, and

the induced dipole in the volume of dielectric medium displaced by atom j. In this model it is assumed that the solute atoms only displace the dielectric medium within their van der Waals radii and therefore does not include cavities within the solute that are too small for solvent molecules. This process is carried out for each atom and although the model is conceptually simple it provides a fast and fairly accurate method for calculating α_i and hence G_{polar}.

3.3 Free Energy of Binding

3.3.1 Definition

The free energy of each of the molecular species in aqueous solution is defined as:

$$\Delta G_{\text{aq}} = E_{\text{gas}} + G_{\text{solvation}} - TS, \tag{3.21}$$

where T is the temperature, S is the entropy, E_{gas} is defined in equation 3.10 and $G_{\text{solvation}}$ is defined in equation 3.12.

For a complex between a protein and a peptide the free energy of binding, $\Delta G_{\text{binding}}$, is defined as:

$$\Delta G_{\text{binding}} = \Delta G_{\text{aq,complex}} - (\Delta G_{\text{aq,protein}} + \Delta G_{\text{aq,peptide}}). \tag{3.22}$$

The equilibrium constant, K, for the non-covalent association of two species, A and B:

$$\text{A} + \text{B} \rightleftharpoons \text{AB}, \tag{3.23}$$

is given by:

$$K = \frac{[\text{AB}]}{[\text{A}]\,[\text{B}]}. \tag{3.24}$$

This can be related to the free energy of binding using the expression:

$$\Delta G_{\text{binding}} = -RT \ln K, \tag{3.25}$$

which allows the concentrations of products and reactants, in particular the IC_{50}, to be explicitly related to changes in free energy.

3.4 Entropy

3.4.1 Introduction

The entropy of a molecule can be estimated, in terms of contributions from translational, rotational and vibrational motions, using classical ideal gas thermodynamics. One of the main ideal gas assumptions is that the particles are not interacting with one another, which is clearly not the case for biomolecular systems. This will introduce errors in the determination of the partition functions. However, it is important to consider the entropic cost of protein–protein association and this method allows quantitative estimates to be made.

3.4.2 Classical Partition Functions and Entropy

Using classical ideal gas thermodynamics, the contribution to the total entropy of a molecule due to any individual component is related to the partition function for that component. The translational contribution to the entropy, S_{trans}, is given by:[101]

$$S_{\text{trans}} = k \left(\ln(z_{\text{trans}}) + \ln(e) + T \left(\frac{\partial \ln(z_{\text{trans}})}{\partial T} \right)_V \right), \tag{3.26}$$

where z_{trans} is the translational partition function, k is the Boltzmann constant, T is the temperature, the subscript V implies constant volume and molar quantites are used. The factor of e arises from the derivation of the expression using Stirling's approximation for the logarithm of a large number.[207] The translational partition function is given by:[101]

$$z_{\text{trans}} = \left(\frac{2\pi mkT}{h^2}\right)^{3/2} \frac{kT}{P}, \tag{3.27}$$

where m is the mass of the molecule, P is the pressure and h is the Planck constant. Since, at constant volume, the ratio $\frac{kT}{P}$ is constant, the partial derivative is:

$$\left(\frac{\partial \ln(z_{\text{trans}})}{\partial T}\right)_V = \frac{3}{2T}, \tag{3.28}$$

and the translational contribution to the entropy, TS_{trans}, is given by:

$$TS_{\text{trans}} = \left[\ln(z_{\text{trans}}) + \frac{5}{2}\right] kT. \tag{3.29}$$

The rotational contribution to the entropy, S_{rot}, is given by:[101]

$$S_{\text{rot}} = k\left(\ln(z_{\text{rot}}) + T\left(\frac{\partial \ln(z_{\text{rot}})}{\partial T}\right)_V\right), \tag{3.30}$$

where z_{rot} is the rotational partition function. For the general case of a non-linear polyatomic molecule, the rotational partition function is given by:[101]

$$z_{\text{rot}} = \frac{1}{\sigma}\left(\frac{8\pi^2 kT}{h^2}\right)^{3/2} (\pi I_{\text{a}} I_{\text{b}} I_{\text{c}})^{1/2}, \tag{3.31}$$

where I_{a}, I_{b} and I_{c} are the moments of inertia of the molecule about its three principal axes and σ is the symmetry number, which is equal to one for non-symmetrical biomolecules. As for the translational partition function, the partial

derivative at constant volume is:

$$\left(\frac{\partial \ln(z_{\text{rot}})}{\partial T}\right)_V = \frac{3}{2T}, \tag{3.32}$$

and the rotational contribution to the entropy, TS_{rot}, is given by:

$$TS_{\text{rot}} = \left[\ln(z_{\text{rot}}) + \frac{3}{2}\right] kT. \tag{3.33}$$

The vibrational contributions to the entropy, TS_{vib}, can be obtained from the normal mode frequencies, ν_j, as follows:[101]

$$TS_{\text{vib}} = \sum_{j=1}^{3N-6} \frac{h\nu_j}{e^{h\nu_j/kT} - 1} - kT \ln\left(1 - e^{-h\nu_j/kT}\right). \tag{3.34}$$

3.5 Energy Minimisation

3.5.1 Minimisation Techniques

There are several techniques for finding the minimum energy geometries of large molecules.[63–65,210] Line searching is a simple method that involves calculating the gradient of the potential energy surface and performing a one-dimensional minimisation in the direction of the line of steepest negative slope. This can be achieved in the TINKER[211] package using the MINIMIZE program, which stores information from previous line searches in order to make the calculations more efficient. This type of minimisation technique is efficient for structures that are far from the minimum but becomes less useful near the minimum. Truncated Newton techniques such as those used in the NEWTON program in TINKER[211] use the gradient and the curvature of the potential energy surface. Approximations are made for the curvature in order to reduce the computational cost of the

calculations. Consequently, these techniques, when used together, are useful for finding minimum energy geometries of large molecules such as proteins or nucleic acids.

3.5.2 The Hessian Matrix

The potential energy of a system, V, depends on p variables such that:

$$V = V(x_1, x_2, \ldots, x_p), \tag{3.35}$$

where the variables $x_i \ldots x_p$ could be bond lengths, bond angles or the cartesian coordinates of the atoms in the molecule. A vector $\mathbf{x}$ can be defined in p-dimensional space as:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}. \tag{3.36}$$

Therefore the Taylor expansion[207] of V about the point x_0 can be written as:[64]

$$\begin{aligned} V(\mathbf{x}) &= V(\mathbf{x}_0) + \sum_{i=1}^{p} \left(\frac{\partial V}{\partial x_i}\right)_{x_i^0} (x_i - x_i^0) + \\ &\quad + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)_{x_i^0 x_j^0} (x_i - x_i^0)(x_j - x_j^0) + \cdots . \end{aligned} \tag{3.37}$$

It is then possible to define the gradient vector $\mathbf{g}$ of the function V as in equation 3.38. When evaluated at an atomic position the gradient of V gives the negative

of the force acting on the atom.

$$\mathbf{g} = \begin{pmatrix} \frac{\partial V}{\partial x_1} \\ \frac{\partial V}{\partial x_2} \\ \vdots \\ \frac{\partial V}{\partial x_p} \end{pmatrix} \tag{3.38}$$

Furthermore, the Hessian or force constant matrix can be defined as follows:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial x_p} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \cdots & \frac{\partial^2 V}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x_p \partial x_1} & \frac{\partial^2 V}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_p^2} \end{pmatrix} . \tag{3.39}$$

At the stationary points of the potential energy function the gradient vector, $\mathbf{g}$, is zero. The nature of the stationary point is determined by the eigenvalues of the Hessian matrix, $\mathbf{H}$. If all of the Hessian eigenvalues are positive or zero then it is a local minimum and if all of the eigenvalues are negative it is a local maximum. If n eigenvalues are negative then it is described as an nth order saddle point.

3.5.3 Normal Modes

For a system of N atoms with masses m_i and position vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$, where the atoms each vibrate about their equilibrium positions, $\mathbf{r}_1^e, \mathbf{r}_2^e, \ldots, \mathbf{r}_N^e$, the mass-weighted displacement coordinates, $\mathbf{Q_i}$, are defined as follows:

$$\mathbf{Q}_i = \sqrt{m_i}\,(\mathbf{r}_i - \mathbf{r}_i^{eq}) . \tag{3.40}$$

A set of $3N$ mass-weighted displacement coordinates $q_1, q_2, \ldots, q_{3N}$ can be defined such that q_1, q_2 and q_3 are the components of $\mathbf{Q}_1$ and so on. As before, the

potential energy, V, of the system can be expressed as a Taylor expansion about the equilibrium position, as follows:

$$V = V_{\text{eq}} + \sum_{i=1}^{3N} \left(\frac{\partial V}{\partial q_i}\right)_{\text{eq}} q_i + \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)_{\text{eq}} q_i q_j + \cdots . \tag{3.41}$$

If the vibrations are small the harmonic approximation can be made and higher order terms in the expansion can be neglected. Since the gradient is zero at equilibrium, the potential energy can be written as follows:

$$V = V_{\text{eq}} + \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)_{\text{eq}} q_i q_j. \tag{3.42}$$

As stated previously the negative of the gradient evaluated at an atomic position gives the force, $\mathbf{F}_i$, acting on atom i. This can be expressed as:

$$-\nabla_i V = \mathbf{F}_i = m_i \frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2}, \tag{3.43}$$

where t is time and the second derivative of position with respect to time is the acceleration on atom i. Hence, each coordinate q must satisfy the following expression:

$$\frac{\mathrm{d}^2 q_i}{\mathrm{d}t^2} + \frac{\partial V}{\partial q_i} = 0. \tag{3.44}$$

If we write $q_i = A_i \sin(\omega t + \phi_i)$ then these terms can be written as in equations 3.45 and 3.46:

$$\frac{\partial V}{\partial q_i} = \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)_{\text{eq}} q_j \quad \text{for } j = 1, 2, \ldots, 3N, \tag{3.45}$$

$$\frac{\mathrm{d}^2 q_i}{\mathrm{d}t^2} = -\omega^2 q_i. \tag{3.46}$$

The sum of the terms in equations 3.45 and 3.46 is zero and this can be written

in matrix form as:

$$\begin{pmatrix} \frac{\partial^2 V}{\partial q_1^2} & \frac{\partial^2 V}{\partial q_1 \partial q_2} & \cdots & \frac{\partial^2 V}{\partial q_1 \partial q_{3N}} \\ \frac{\partial^2 V}{\partial q_2 \partial q_1} & \frac{\partial^2 V}{\partial q_2^2} & \cdots & \frac{\partial^2 V}{\partial q_2 \partial q_{3N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial q_{3N} \partial q_1} & \frac{\partial^2 V}{\partial q_{3N} \partial q_2} & \cdots & \frac{\partial^2 V}{\partial q_{3N}^2} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_{3N} \end{pmatrix} = \omega^2 \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_{3N} \end{pmatrix}. \tag{3.47}$$

The matrix on the left is the mass-weighted Hessian matrix and its eigenvalues give the vibrational frequencies of the normal modes of vibration. For a non-linear molecule there are six zero eigenvalues, corresponding to three translational and three rotational normal modes. The eigenvalues, ϵ, are found by solving the following determinant:

$$\begin{vmatrix} k_{11} - \epsilon & k_{12} & \cdots & k_{13N} \\ k_{21} & k_{22} - \epsilon & \cdots & k_{23N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{3N1} & k_{3N2} & \cdots & k_{3N3N} - \epsilon \end{vmatrix} = 0, \tag{3.48}$$

where $k_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j}$. These eigenvalues can then be used to determine the eigenfunctions, which give the relationship between the normal coordinates and the original Cartesian coordinates.

3.6 Molecular Dynamics

3.6.1 Introduction

Molecular dynamics (MD) simulations are among the principal tools used in the study of biomolecules as they allow the time-dependent behaviour of a molecular system to be determined.[72] The trajectory describing the motion of each

particle within a system can be determined by integrating Newton's second law, also called the equation of motion. The force acting on each atom is calculated from the potential energy surface. The trajectory can then be studied using statistical techniques to calculate average values of a wide range of properties. MD simulations are deterministic in the sense that once the positions and velocities of the atoms are known, the positions and velocities can be calculated, in principle, for any value of time.

3.6.2 Newton's Equations of Motion

Newton's equation of motion can be expressed as:

$$\mathbf{F}_i = m_i \mathbf{a}_i = -\nabla_i V, \tag{3.49}$$

where $\mathbf{F}_i$ is the force on particle i, m_i is the mass of particle i, $\mathbf{a}_i$ is the acceleration of particle i and V is the potential energy. This relation can also be expressed as:

$$-\frac{\mathrm{d}V}{\mathrm{d}\mathbf{r}_i} = m_i \frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2}, \tag{3.50}$$

allowing the gradient of the potential energy of the system to be related to the changes of atomic positions, $\mathbf{r}_i$, as a function of time.

For the one-dimensional case where acceleration is constant and position is given by $\mathbf{x}$:

$$\mathbf{a} = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}, \tag{3.51}$$

integration gives:

$$\mathbf{v} = \mathbf{a}t + v_0, \tag{3.52}$$

and since:

$$\mathbf{v} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}, \tag{3.53}$$

it also follows by integration that:

$$\mathbf{x} = \mathbf{v}t + x_0. \tag{3.54}$$

Thus, the value of $\mathbf{x}$ at time t is given by:

$$\mathbf{x} = \mathbf{a}t^2 + v_0 t + x_0, \tag{3.55}$$

where v_0 is the initial velocity and x_0 is the initial position.

Therefore, all that is needed to calculate a trajectory are the initial atomic positions, a distribution of initial velocities and the acceleration, which is determined by the gradient of the potential energy function. This allows the position, velocity and acceleration for each atom within a system to be determined at any given value of time, t.

The initial positions can be taken from experimental data and the initial distribution of velocities is usually determined randomly from a Maxwell-Boltzmann or Gaussian distribution. The velocity distribution is chosen to correspond to the required system temperature, T, and such that there is no overall momentum.

3.6.3 Integration Algorithms

The potential energy is a function of $3N$ variables for N atoms. Due to its complicated nature, there is no analytical solution and the equations must be solved numerically. Taylor expansions[207] can be used to approximate the

positions, velocities and accelerations of the atoms:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t)\,\delta t + \frac{1}{2}\mathbf{a}(t)\,\delta t^2 + \cdots \tag{3.56}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \mathbf{a}(t)\,\delta t + \frac{1}{2}\mathbf{b}(t)\,\delta t^2 + \cdots \tag{3.57}$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \mathbf{b}(t)\,\delta t + \cdots, \tag{3.58}$$

where $\mathbf{r}$ is the position, $\mathbf{v}$ is the velocity, $\mathbf{a}$ is the acceleration and $\mathbf{b}$ is the third derivative of $\mathbf{r}$ with respect to t.

Some of the commonly used integration algorithms are given below.[72] All of the algorithms described here are essentially equivalent and generate identical position trajectories. Their main differences occur in the treatment of the velocity propagation and how many variables need to be stored to calculate a trajectory. Other methods, such as the predictor-corrector method of Gear,[212,213] can also be used to propagate MD trajectories but are not described further here.

3.6.3.1 Verlet Algorithm

The Verlet algorithm[214] is one of the simplest methods of solving the equations of motion. It is based on the positions and accelerations at time t and the positions at time $t - \delta t$. The Taylor expansion for the positions at time $t + \delta t$ is given in equation 3.56. The Taylor expansion for the positions at time $t - \delta t$ is given by:

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \mathbf{v}(t))\delta t + \frac{1}{2}\mathbf{a}(t)\,\delta t^2 + \cdots, \tag{3.59}$$

and adding these two expressions together gives the expression for the propagation of the positions:

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2 \mathbf{a}(t). \tag{3.60}$$

This expression does not explicitly contain the velocities but they can be calculated from:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t}, \tag{3.61}$$

although this expression is not exact. More accurate expressions for $\mathbf{v}(t)$ require the storage of more variables and increase the computational cost of the calculations, as well as requiring the positions $\mathbf{r}(t+\delta t)$ to be known.

3.6.3.2 Leap-Frog Algorithm

The leap-frog algorithm[215] treats the velocities explicitly by calculating them at half-step intervals. The velocities are calculated at time $t + \frac{1}{2}\delta t$ and used to calculate the positions in the following scheme:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}\left(t+\frac{1}{2}\delta t\right)\delta t \tag{3.62}$$

$$\mathbf{v}\left(t+\frac{1}{2}\delta t\right) = \mathbf{v}\left(t-\frac{1}{2}\delta t\right) + \mathbf{a}(t)\,\delta t. \tag{3.63}$$

Equation 3.63 is implemented first and the velocities are said to leap over the positions to the next half-step velocities. The velocities at time t can be evaluated during this stage using:

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}\left(t+\frac{1}{2}\delta t\right) + \mathbf{v}\left(t-\frac{1}{2}\delta t\right)\right], \tag{3.64}$$

giving the positions and velocities at the same instant. Equation 3.62 is then used to propagate the positions, which again leap over the velocities, before the accelerations are determined for the next step.

3.6.3.3 Velocity Verlet Algorithm

The velocity Verlet algorithm[216] is equivalent to the Verlet algorithm[214] but treats the velocities explicitly. The algorithm takes the form:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t)\,\delta t + \frac{1}{2}\mathbf{a}(t)\,\delta t^2 \tag{3.65}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}\left[\mathbf{a}(t) + \mathbf{a}(t+\delta t)\right]\delta t. \tag{3.66}$$

The positions at time $t + \delta t$ are calculated first and half-step velocities are determined using:

$$\mathbf{v}\left(t + \frac{1}{2}\delta t\right) = \mathbf{v}(t) + \frac{1}{2}\mathbf{a}(t)\,\delta t. \tag{3.67}$$

The forces and, hence, the accelerations at time $t + \delta t$ are evaluated and the velocity propagation is completed:

$$\mathbf{v}(t+\delta t) = \mathbf{v}\left(t + \frac{1}{2}\delta t\right) + \frac{1}{2}\mathbf{a}(t+\delta t)\,\delta t. \tag{3.68}$$

This method also requires the storage of just three variables: the positions, velocities and accelerations at time t.

3.6.3.4 Beeman Algorithm

The Beeman algorithm[217] is similar to the velocity Verlet algorithm[216] but uses the following expressions for the positions and velocities:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t)\,\delta t + \frac{2}{3}\mathbf{a}(t)\,\delta t^2 - \frac{1}{6}\mathbf{a}(t-\delta t)\,\delta t^2 \tag{3.69}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \mathbf{v}(t)\,\delta t + \frac{1}{3}\mathbf{a}(t+\delta t)\,\delta t + \frac{5}{6}\mathbf{a}(t)\,\delta t + -\frac{1}{6}\mathbf{a}(t-\delta t)\,\delta t. \tag{3.70}$$

This is the integration algorithm used in the molecular dynamics simulations described in this thesis. This method requires the storage of four variables: the

positions, velocities and accelerations at time t, as for velocity Verlet, as well as the accelerations at time $t - \delta t$. However, the treatment of the velocities is improved over the standard Verlet method.[72]

3.6.4 Long Range Interactions

3.6.4.1 Periodic Boundary Conditions

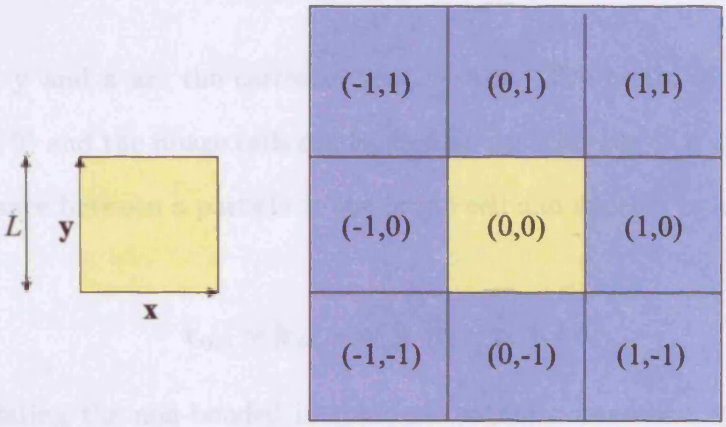

Figure 3.1: (Left) The unit cell and its coordinates, **x** and **y**, for a square box of length L in 2D. (Right) A 3x3 periodic lattice built from unit cells. Each cell's coordinate vector is given, with the original cell located at (0,0). Figure adapted from Toukmaji and Board.[218]

In the periodic boundary conditions (PBC) approach, the simulation box is infinitely replicated in each direction to form a lattice of identical boxes. The simulation box can take any shape that will fill space as it is periodically replicated, although the case of a cubic box of length L will be considered here. The origin cell and the first layer of replicas for a 2-dimensional system are shown in figure 3.1. The array of boxes in 3-dimensional space is conceptually built up in approximately spherical layers starting from the initial simulation box. The location of a box within the infinite lattice is described in terms of a coordinate vector, defined as:

$$\mathbf{n} = (n_1, n_2, n_3) = n_1 L\mathbf{x} + n_2 L\mathbf{y} + n_3 L\mathbf{z}, \tag{3.71}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are the cartesian unit vectors. The origin cell is located at $\mathbf{n} = (0, 0, 0)$ and the image cells are located at $L\mathbf{n}$ intervals as $\mathbf{n}$ goes to infinity. The distance between a particle in the origin cell and another in an image cell is given by:

$$\mathbf{r}_{ij,\mathbf{n}} = |\mathbf{r}_{j\mathbf{n}} - \mathbf{r}_i| = |\mathbf{r}_i - \mathbf{r}_j + L\mathbf{n}|\,. \tag{3.72}$$

Calculating the non-bonded interactions within a molecular system involves determining interactions between all pairs of atoms. For a large system, for example a protein in solution, this can be a huge number of interactions and this significantly increases the computational cost of a simulation. Consequently, it is common to use more approximate methods such as cutoff approaches and Ewald summation to deal with the van der Waals and electrostatic interactions. The treatment of these interactions is particularly important for the electrostatic terms as these are effective over a much longer range, varying as the recriprocal of

the interatomic separation. In particular, the use of PBC introduces the problem of determining interactions between a molecule and the infinite replicas of itself.

3.6.4.2 Cutoff Methods

Cutoff approaches can be used for any non-bonded interactions but are more commonly used for shorter range van der Waals interactions. The approach is to set all interactions involving atoms separated by more than some cutoff distance, r_{cutoff}, to zero. The simplest way to achieve this is by the use of a truncation function such as:[68]

$$S(r) = \begin{cases} 1 & r \leq r_c \\ 0 & r > r_c \end{cases}, \tag{3.73}$$

which is multiplied by each term of the non-bonded interaction energy. This achieves the desired result in terms of reducing the cost of the energy calculation but has the drawback that the energy is no longer a continuous function and its derivatives, from which the forces on atoms are determined, also become discontinuous. This can cause problems in energy minimisations or MD simulations as atoms pass in and out of each other's cutoff distance. There are two general approaches to avoiding this problem: switch functions and shift functions. A switch function such as:[68]

$$S(r) = \begin{cases} 1 & r \leq r_{\text{on}} \\ \frac{\left(r_{\text{cutoff}}^2 - r^2\right)^2 \left(r_{\text{cutoff}}^2 + 2r^2 - 3r_{\text{on}}^2\right)}{\left(r_{\text{cutoff}}^2 - r_{\text{on}}^2\right)^3} & r_{\text{on}} < r \leq r_{\text{cutoff}} \\ 0 & r > r_{\text{cutoff}} \end{cases}, \tag{3.74}$$

does not modify the non-bonded interactions for separations less than an inner cutoff, r_{on}, and smoothly reduces the interactions to zero between r_{on} and r_{cutoff}.

The energy and its first derivative are both continuous throughout the whole range of r. A shift function such as:[68]

$$S(r) = \begin{cases} \left[1 - \left(\frac{r}{r_{\text{cutoff}}}\right)^2\right]^2 & r \leq r_{\text{cutoff}} \\ 0 & r > r_{\text{cutoff}} \end{cases}, \tag{3.75}$$

uses only one cutoff distance and modifies the interaction for all interatomic separations. Once again, the energy and its first derivative are both continuous throughout the whole range of r. Both switch and shift approaches are therefore suitable for dynamics calculations.

The simplest way to apply these methods is to calculate each interaction and apply the truncation function. However, this does not reduce the cost of the calculation as all of the interactions must still be evaluated. This is avoided by keeping a list of all the non-bonded interactions that involve interatomic separations less than the list cutoff, r_{list}, which is larger than r_{cutoff}. This means that more interactions are calculated than are within the cutoff but significantly reduces the number of interactions that must be evaluated. The list is updated when the current list becomes invalid,[68] i.e. the atoms have moved by a similar amount to the difference between the list and interaction cutoffs, $r_{\text{list}} - r_{\text{cutoff}}$.

Within the PBC approach, it is common to restrict the cutoff distance to half the length of the simulation box. This has the effect of ensuring that each particle in the box only interacts with particles in the nearest layer of replicated cells. This is known as the minimum-image convention.[68] Although this method is easy to implement and appropriate for van der Waals interactions, its use for treatment of electrostatic interactions has been shown to introduce errors and artificial behaviour into simulations, meaning that another approach must be

used.[218–221]

3.6.4.3 Particle Mesh Ewald Summation

Ewald summation[222] provides an efficient method for calculating the long-range electrostatic interactions between the particles and all of their infinite periodic images. Particle mesh Ewald (PME) summation[223] is an improvement over standard Ewald techniques but follows the same basic recipe. The total electrostatic energy of a system of N particles in a cubic box of length L and their infinite replicas within PBC is given by:[218]

$$E_{\text{electrostatic}}^{\text{total}} = \frac{1}{2}\sum_{\mathbf{n}}{}' \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{\mathbf{r}_{ij,\mathbf{n}}}, \tag{3.76}$$

where q_i is the charge on particle i and n is the cell coordinate vector, defined previously. The prime on the first sum indicates that terms are omitted when $\mathbf{n} = \mathbf{0}$.

In order to decompose this sum, consider that each point charge in the sum is surrounded by a Gaussian charge distribution of equal magnitude and opposite sign. This has the effect of screening the charges from one another, effectively making the interactions short-ranged, which causes the sum in real space to converge rapidly. In order to counteract the induced charges, a second Gaussian distribution is added with the same magnitude and sign as the original point charges. The sum is then performed in reciprocal space using fast Fourier transform techniques. The electrostatic energy can thus be re-written:[222]

$$E_{\text{Ewald}} = E_{\text{real}} + E_{\text{reciprocal}} + E_{\text{self}} + E_{\text{surface}}, \tag{3.77}$$

where E_{real} is the real space or direct sum, $E_{\text{reciprocal}}$ is the reciprocal, imaginary

or Fourier sum, E_{self} is the self term and E_{surface} term. The real sum is given by:

$$E_{\text{real}} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\mathbf{n}=0}^{\infty}{}' q_i q_j \frac{\operatorname{erfc}(\alpha \mathbf{r}_{ij,\mathbf{n}})}{\mathbf{r}_{ij,\mathbf{n}}}, \tag{3.78}$$

where α is a positive parameter that determines the width of the induced Gaussian distribution and erfc is the complementary error function, given by:

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x \exp\left(-u^2\right) \mathrm{d}u, \tag{3.79}$$

which rapidly falls to zero as interatomic separation increases. The reciprocal space term is given by:

$$E_{\text{reciprocal}} = \frac{1}{2\pi V}\sum_{\mathbf{m}\neq 0} \frac{\exp\left(-\left(\frac{\pi \mathbf{m}}{\alpha}\right)^2\right)}{\mathbf{m}^2} S(\mathbf{m}) S(-\mathbf{m}), \tag{3.80}$$

where V is the volume of the simulation box, $\mathbf{m}$ is a reciprocal space vector and $S(\mathbf{m})$ is the structure factor, defined as:[218]

$$S(\mathbf{m}) = \sum_{k=1}^{N} q_k \exp(2\pi i \mathbf{m} \cdot \mathbf{r}_k). \tag{3.81}$$

The self term corrects for the interactions of each of the induced charges with itself and is given by;

$$E_{\text{self}} = -\frac{\alpha}{\sqrt{\pi}}\sum_{i=1}^{N} q_i^2. \tag{3.82}$$

The surface term arises due to the interaction of a dipole on the (roughly spherical) surface of the array of periodic boxes and the surrounding medium.[224] For an infinite dielectric constant in the surrounding medium, this term is zero and tin-foil boundary conditions are said to be employed. For a surrounding vacuum with a dielectric constant of 1, the surface term is given by:[72]

$$E_{\text{surface}} = \frac{2\pi}{3V}\left|\sum_{i=1}^{N} q_i \mathbf{r}_i\right|^2 \tag{3.83}$$

In PME,[223] the real space sum is calculated using cutoffs while the reciprocal sum is approximated using fast Fourier transforms (FFT) on a grid where charges are interpolated to grid points.[218] The parameter α is chosen to be large enough that a fixed cutoff can be used for the direct sum, reducing its formal scaling from N^2 to N. The scaling of the FFT approach used is formally $\log(N)$ giving PME a scaling of $N\log(N)$.[218,223]

Ewald approaches employ a neutralising background plasma for systems with a non-zero net charge. This can lead to artificial behaviour, particularly where the volume of the box changes, such as in constant pressure MD.[225] This can be avoided by the use of counterions to neutralise the system electronically. Alternatively, a correction term such as that proposed by Bogusz *et al.*[225] can be introduced to reduce any artificial effects.

3.6.5 Constraints

3.6.5.1 Geometric Constraints

It is often desirable to constrain one or more of the degrees of freedom within an MD simulation. In particular, constraint of high frequency motions such as bond vibrations can allow the use of a longer time step. The use of longer time steps without constraining these high frequency motions can lead to problems as these motions occur on a similar timescale to the time step. This is particularly pronounced for bonds involving hydrogen atoms.[1]

A general approach to constraining bond lengths in MD simulations[226] involves artificially introducing forces acting along the bonds. A normal MD time step

is calculated without the constraints and the magnitude of the constraint forces are determined such that they correct the atomic positions and keep the bond lengths constant. The equation of motion, equation 3.49, becomes:

$$m_i \mathbf{a}_i = \mathbf{F}_i + \mathbf{G}_i \approx \mathbf{F}_i + \mathbf{G}_i^a, \tag{3.84}$$

where $\mathbf{G}_i$ are the constraint forces acting on each atom i and the other terms are as before. In practice, since the equations of motion are integrated numerically, the constraint forces are generally evaluated to ensure that the constraints are satisfied for each time step.[72] Thus, approximate restraint forces, $\mathbf{G}_i^a$, are calculated for each atom.[226]

This method can be illustrated by considering the example of restraining the bond lengths in water, H_2O, where the oxygen atom is labelled atom 2 and the hydrogens are labelled atoms 1 and 3, respectively.[72] The equations of motion for the three atoms in the water molecule are:

$$m_1 \mathbf{a}_1 = \mathbf{F}_1 + \mathbf{G}_1 \approx \mathbf{F}_1 + \mathbf{G}_1^a \tag{3.85}$$

$$m_2 \mathbf{a}_2 = \mathbf{F}_2 + \mathbf{G}_2 \approx \mathbf{F}_2 + \mathbf{G}_2^a \tag{3.86}$$

$$m_3 \mathbf{a}_3 = \mathbf{F}_3 + \mathbf{G}_3 \approx \mathbf{F}_3 + \mathbf{G}_3^a, \tag{3.87}$$

and the constraint forces, $\mathbf{G}_i$, maintain the following constraints:

$$\chi_{12} = \mathbf{r}_{12}^2(t) - d_{12}^2 = 0 \tag{3.88}$$

$$\chi_{23} = \mathbf{r}_{23}^2(t) - d_{23}^2 = 0, \tag{3.89}$$

where $\mathbf{r}_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ and d_{12} and d_{23} are the bond lengths. Within the context of the Verlet algorithm,[214] the propagation of the positions can be written as:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i'(t + \delta t) + \left(\frac{\delta t^2}{m_i}\right) \mathbf{G}_i^a(t), \tag{3.90}$$

where $\mathbf{r}_i'(t+\delta t)$ represents the position that would have been reached after an unrestrained time step. For the example of water, the approximations of the restraint forces are:

$$\mathbf{G}_1^a = \lambda_{12}\mathbf{r}_{12} \tag{3.91}$$

$$\mathbf{G}_2^a = \lambda_{23}\mathbf{r}_{23} - \lambda_{12}\mathbf{r}_{12} \tag{3.92}$$

$$\mathbf{G}_3^a = -\lambda_{23}\mathbf{r}_{23}, \tag{3.93}$$

where λ_{12} and λ_{23} are the multipliers that are yet to be determined. These forces act along the directions of the bonds and obey Newton's third law.[72] Writing the Verlet positions explicitly gives:

$$\mathbf{r}_1(t+\delta t) = \mathbf{r}_1'(t+\delta t) + \left(\frac{\delta t^2}{m_1}\right)\lambda_{12}\mathbf{r}_{12}(t) \tag{3.94}$$

$$\mathbf{r}_2(t+\delta t) = \mathbf{r}_2'(t+\delta t) + \left(\frac{\delta t^2}{m_2}\right)\lambda_{23}\mathbf{r}_{23} - \left(\frac{\delta t^2}{m_2}\right)\lambda_{12}\mathbf{r}_{12}(t) \tag{3.95}$$

$$\mathbf{r}_3(t+\delta t) = \mathbf{r}_3'(t+\delta t) - \left(\frac{\delta t^2}{m_3}\right)\lambda_{23}\mathbf{r}_{23}(t). \tag{3.96}$$

Therefore, we can also write:

$$\mathbf{r}_{12}(t+\delta t) = \mathbf{r}_{12}'(t+\delta t) + \left(\frac{\delta t^2}{m_1+m_2}\right)\lambda_{12}\mathbf{r}_{12}(t) - \left(\frac{\delta t^2}{m_2}\right)\lambda_{23}\mathbf{r}_{23}(t) \tag{3.97}$$

$$\mathbf{r}_{23}(t+\delta t) = \mathbf{r}_{23}'(t+\delta t) - \left(\frac{\delta t^2}{m_2}\right)\lambda_{12}\mathbf{r}_{12}(t) + \left(\frac{\delta t^2}{m_2+m_3}\right)\lambda_{23}\mathbf{r}_{23}(t). \tag{3.98}$$

The next step is to take the square modulus of both sides and apply the constraints:

$$|\mathbf{r}_{12}(t+\delta t)|^2 = |\mathbf{r}_{12}(t)|^2 = d_{12}^2 \tag{3.99}$$

$$|\mathbf{r}_{23}(t+\delta t)|^2 = |\mathbf{r}_{23}(t)|^2 = d_{23}^2, \tag{3.100}$$

which yields a pair of quadratic equations in λ_{12} and λ_{23}. These equations are solved iteratively to determine the multipliers, which are then used to correct

the atomic positions. This method can also be extended to include bond angle constraints, for example by fixing the H–H separation in the water molecule.[72]

For large molecules such as proteins, the number of constraints is significant and these constraints have to be evaluated cyclically. Each constraint is calculated individually for one iteration and the process is repeated until all of the constraints are satisfied, to within some designated level of tolerance. This approach is referred to as the SHAKE algorithm[226,227] and is most easily applied to the Verlet integration method,[72] although it has been implemented in other integration schemes.[228] A similar algorithm, implemented in explicit-velocity schemes such as the velocity Verlet or Beeman algorithms, is called RATTLE.[229] For example, in the velocity Verlet scheme, as described in section 3.6.3.3, the first stage of the integration involves calculating the positions at time $t + \delta t$ and the half-step velocities with the constraint forces:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}'_i(t+\delta t) + \frac{1}{2}\left(\frac{\delta t^2}{m_i}\right)\mathbf{G}_i^a(t) \tag{3.101}$$

$$\mathbf{v}_i\left(t+\frac{1}{2}\delta t\right) = \mathbf{v}'_i\left(t+\frac{1}{2}\delta t\right) + \frac{1}{2}\left(\frac{\delta t}{m_i}\right)\mathbf{G}_i^a(t), \tag{3.102}$$

and the second stage involves completing the velocity propagation using:

$$\mathbf{v}_i(t+\delta t) = \mathbf{v}'_i(t+\delta t) + \frac{1}{2}\left(\frac{\delta t}{m_i}\right)\mathbf{G}_i^a(t+\delta t). \tag{3.103}$$

RATTLE[229] is the algorithm used in this work to constrain the bond lengths during the MD simulations.

3.6.5.2 Temperature and Pressure Constraints

Berendsen *et al.*[230] described a method for performing MD with coupling to an external bath at constant temperature and pressure. The temperature coupling

essentially involves scaling the atomic velocities at each time step from $\mathbf{v}$ to $\lambda\mathbf{v}$, where λ is given by:

$$\lambda = 1 + \frac{\delta t}{2\tau_{\mathrm{T}}}\left(\frac{T_0}{T} - 1\right), \tag{3.104}$$

where τ_{T} is the temperature coupling parameter, with dimensions of time, T is the instantaneous temperature and T_0 is the desired reference temperature. The change in temperature per time step can be set as $\frac{(T_0 - T)\delta t}{\tau_{\mathrm{T}}}$, which gives:

$$\lambda = \left[1 + \frac{\delta t}{\tau_{\mathrm{T}}}\left(\frac{T_0}{T} - 1\right)\right]^{\frac{1}{2}}. \tag{3.105}$$

The temperature coupling parameter, τ_{T}, controls the strength of the coupling. If τ_{T} is large, the coupling is weak and the temperature is perturbed slowly by the velocity scaling and *vice versa.* This perturbation acts like a frictional force, reducing the temperature if the system temperature is too high. Similarly, if the system temperature is too low, the coupling scales the velocities such that the temperature increases towards the reference temperature.[68]

A similar approach is used to maintain constant pressure by coupling to an external bath. The scaling is performed on the coordinates and the dimensions, L, of the simulation box: $\mathbf{r}$ scales to $\mu\mathbf{r}$ and L is scaled to μL, where μ is given by:

$$\mu = 1 - \frac{\beta\delta t}{3\tau_{\mathrm{P}}}\left(P_0 - P\right), \tag{3.106}$$

where β is the isothermal compressibility, τ_{P} is the pressure coupling parameter, P is the instantaneous pressure and P_0 is the desired reference pressure. The coupling parameter τ_{P} again has dimensions of time and controls the strength of the pressure coupling in the same way as τ_{T} controls the temperature coupling.[68]

An equivalent expression for μ is:[230]

$$\mu = \left[1 - \frac{\delta t}{\tau_P}\left(P_0 - P\right)\right]^{\frac{1}{3}}. \tag{3.107}$$

This method of coupling to an external bath is the basis of the temperature and pressure controls used in the MD simulations described in this work. This method can be used in conjunction with geometric constraint methods such as SHAKE or RATTLE.[230]

Chapter 4

Methods

4.1 Mdm2 with p53-derived peptides

4.1.1 Amino Acid Sequences

The atomic coordinates of the complex formed between *X. Laevis* Mdm2 (xMdm2) and an 11-residue peptide derived from p53[5] were obtained in Protein Data Bank (PDB) format from the Research Collaboratory for Structural Bioinformatics (RCSB) database.[2] The atomic coordinates were also obtained for the complex formed between human Mdm2 (hMdm2) and a 13-residue p53-derived peptide.[5] These PDB files were converted to TINKER[211] Cartesian coordinate files. Any histidine residues were uncharged to reflect the most likely protonation state at physiological pH. The amino acid sequences for xMdm2, hMdm2 and the p53-derived peptides are given in Table 4.1. The AMBER force field was used with the parm94 parameters of Cornell *et al.*[74]

	xMdm2–p53
Protein	EKLVQPTPLLLSLLKSAGAQKETFTMKEVIYHLGQYIMAKQLYDEKQQHIVHCSNDPLGELFGVQEFSVKEPRRLYAMISRNLVSANV
Peptide	ETFSDLWKLLP
	hMdm2–p53
Protein	ETLVRPKPLLLKLLKSVGAQKDTYTMKEVLFYLGQYIMTKRLYDEKQQHIVYCSNDLLGDLFGVPSFSVKEHRKIYTMIYRNLVV
Peptide	ETFSDLWKLLPEN

Table 4.1: Amino acid sequences for *X. Laevis* and human Mdm2 and the p53-derived peptides used in this work.[5]

Throughout this work, the term peptide is used to describe the smaller of the interacting polypeptides. In this case, the p53-derived peptides consist of the 11-residue (17–27) stretch and the 13-residue (17–29) stretch of human p53, respectively. The term protein is used to describe the larger of the two interacting biomolecules, with xMdm2 consisting of residues 21–108 of the *X. Laevis* enzyme and hMdm2 consisting of residues 25–109 of the human enzyme. The term complex refers to the complex formed between the protein and the peptide in each case. The structure of the xMdm2–p53 complex used in this work is given in Figure 4.1.

4.1.2 Comparison with Other Studies

The Mdm2–p53 system was chosen to test the methodology in its present implementation as an ideal previous study was available for comparison. Massova and Kollman[3] inserted a 12-residue peptide, consisting of residues 16–27 of human p53, into the binding sites of both xMdm2 and hMdm2. The proteins consisted

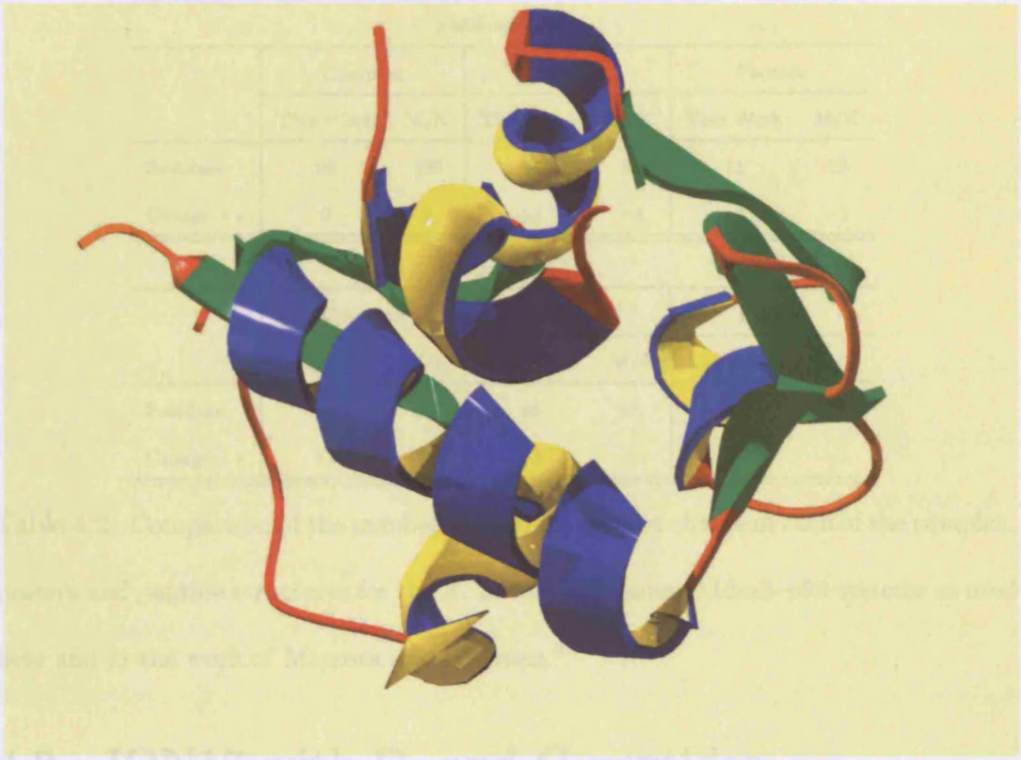

Figure 4.1: Structure of the complex formed between *X. Laevis* Mdm2 and a stretch of the tumour suppressor peptide p53.

of the same residues as those described in this work. The peptide sequence (QETFSDLWKLLP) is similar to those used in this work. The differences in the number of residues and net charge of each of the complex, protein and peptide structures are summarised in Table 4.2.

xMdm2–p53						
	Complex		Protein		Peptide	
	This Work	M/K	This Work	M/K	This Work	M/K
Residues	99	100	88	88	11	12
Charge / *e*	0	0	+1	+1	−1	−1
hMdm2–p53						
	Complex		Protein		Peptide	
	This Work	M/K	This Work	M/K	This Work	M/K
Residues	98	97	85	85	13	12
Charge / *e*	+3	+4	+5	+5	−2	−1

Table 4.2: Comparison of the number of residues and net charge of each of the complex, protein and peptide structures for the *X. Laevis* and human Mdm2–p53 systems as used here and in the work of Massova and Kollman.[3]

4.2 IQN17 with D- and C-peptides

4.2.1 Amino Acid Sequences

The atomic coordinates of the complex formed between IQN17 and D10-p1 were obtained in PDB format, as before. Due to the unnatural, mirror image nature of the D-peptide, it was not possible to convert the PDB file directly into a TINKER Cartesian coordinate file. The PDB file was edited in such a way that the D-amino acids were changed to L-amino acids and the coordinates were inverted through the origin. Thus the conversion to TINKER Cartesian coordinates could be carried out to give an L-peptide version of the desired peptide with coordinates that were inverted from the desired D-peptide. Another inversion through the origin was performed on the TINKER file to give the orig-

IQN17	RMKGIEDKIEEIESKQKKIENEIARIKKLLQLTVWGIKQLQARIL
D10-p1	GACEARHREWAWLCAA
C12	(MT)WEEWDREIENYT

Table 4.3: Amino acid sequences for IQN17, D10-p1[21] and C12.[40] The residues in brackets are those that complete the C14 sequence.

inal complex, retaining the recognition of the amino acid types and hence the appropriate atom types. The AMBER force field was used with the parm94 parameters of Cornell *et al.*[74] The structure of the IQN17–D10-p1 complex used in this work is given in Figure 4.2.

The atomic coordinates of the complex formed between IQN17 and C14linkmid (C14lm) were also obtained in PDB format, as before. A twelve residue stretch of C14 is used throughout this work and referred to as C12. C12lm is bridged at the same residues as C14lm. TINKER could not assign appropriate atom types to the non-peptide diaminoalkane bridging atoms of the IQN17–C12lm and IQN17–C12unlm complexes meaning that no partial charge parameters were available for those atoms. These parameters were assigned by hand by considering similar functional groups that were already parameterised within the TINKER implementation of the AMBER[74] force field and making adjustments based on the electronegativity of each of the atoms. The formal net charge for the peptide was maintained throughout. The assigned partial charges and the atom types that were used as the basis of these assignments are given in Table 4.4. Although these partial charges may differ from their "correct" values to a certain extent, they are not expected to significantly affect the outcome of the simulations, since

Atom	Electronegativity	Partial Charge / e	Atom	Partial Charge / e
N-bound C	2.55	0.3309	Glutamine HE2	0.4251
H of N-bound C	2.20	0.0236	Isoleucine HG1	0.0236
sp^2 bridge C	2.55	−0.0430	Isoleucine CG1	−0.0430
H of sp^2 bridge C	2.20	0.0236	Isoleucine HG1	0.0236
Glutamine N	3.07	−0.9000	Glutamine NE2	−0.9407
sp^2 broken bridge C	2.55	−0.0367	Isoleucine CG1	−0.0430
H of sp^2 broken bridge C	2.20	0.0236	Isoleucine HG1	0.0236
sp^3 broken bridge C	2.55	−0.0600	Isoleucine CD1	−0.0660
H of sp^3 broken bridge C	2.20	0.0186	Isoleucine HD1	0.0186

Table 4.4: Partial charges for non-peptide bridging atoms in C12linkmid and C12unlinkmid. The peptide atom types that were used as the basis of the assignment of these partial carges are also given. The electronegativity values are given according to the Pauling scale.[231]

the bridging atoms are not directly involved in the binding interface. Clearly, such a method of assigning partial charges by hand should only be used for a small number of atoms in a system. The total number of atoms, residues and the net charge of each of the complex, protein and peptide systems are given in Table 4.5. The structure of the IQN17–C12lm complex is given in Figure 4.3.

4.2.2 The IQN17 Hydrophobic Pocket

As discussed in Chapter 1, the fusogenic conformation of gp41 consists of C-terminal helices bound to the trimeric N-terminal core along hydrophobic grooves with hydrophobic pockets at one end.[12,21,25,35] Each of the three hydrophobic

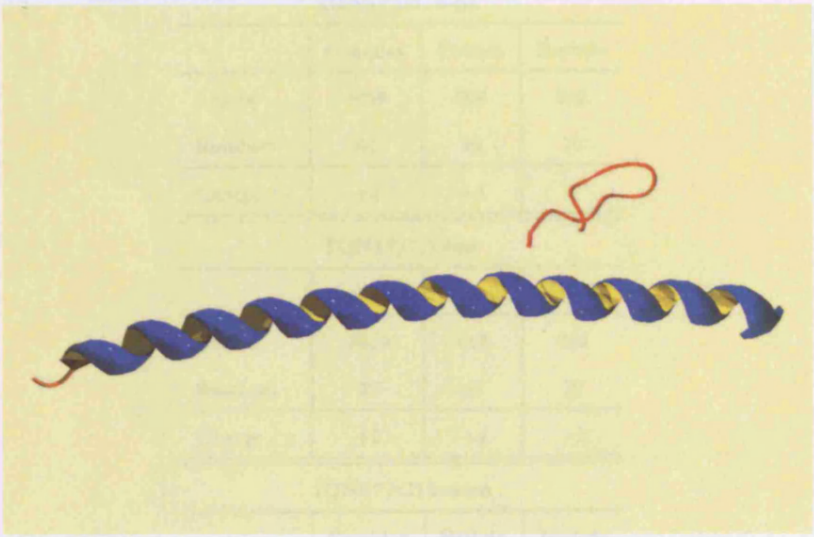

Figure 4.2: Structure of the complex formed between one helix of IQN17 and D10-p1.

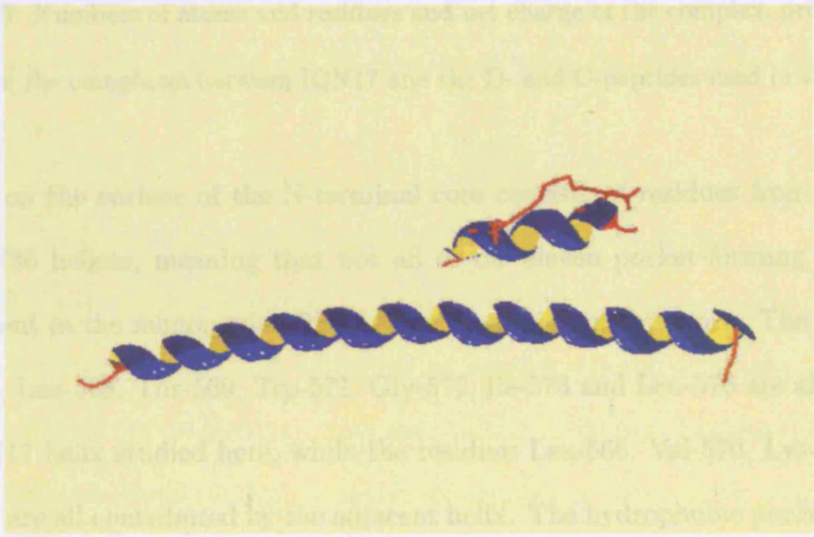

Figure 4.3: Structure of the complex formed between one IQN17 helix and C12linkmid (C12lm).[40] The diaminoalkane linker is shown in red.

IQN17/D10-p1			
	Complex	Protein	Peptide
Atoms	1050	808	242
Residues	61	45	16
Charge / e	+4	+4	0
IQN17/C12lm			
	Complex	Protein	Peptide
Atoms	1040	808	232
Residues	57	45	12
Charge / e	+2	+4	−2
IQN17/C12unlm			
	Complex	Protein	Peptide
Atoms	1045	808	237
Residues	57	45	12
Charge / e	+2	+4	−2

Table 4.5: Numbers of atoms and residues and net charge of the complex, protein and peptide in the complexes between IQN17 and the D- and C-peptides used in this work.

pockets on the surface of the N-terminal core consists of residues from two adjacent N36 helices, meaning that not all of the eleven pocket-forming residues are present in the monomeric IQN17 systems studied in this work. The residues Leu-565, Leu-568, Thr-569, Trp-571, Gly-572, Ile-573 and Leu-576 are all part of the IQN17 helix studied here, while the residues Leu-566, Val-570, Lys-574 and Gln-577 are all contributed by the adjacent helix. The hydrophobic pocket region of the IQN17–D10-p1 complex is shown in detail in Figure 4.4, illustrating the positions of the missing pocket residues. The pocket region of the IQN17–C12lm complex is given in Figure 4.5, which also demonstrates the different binding

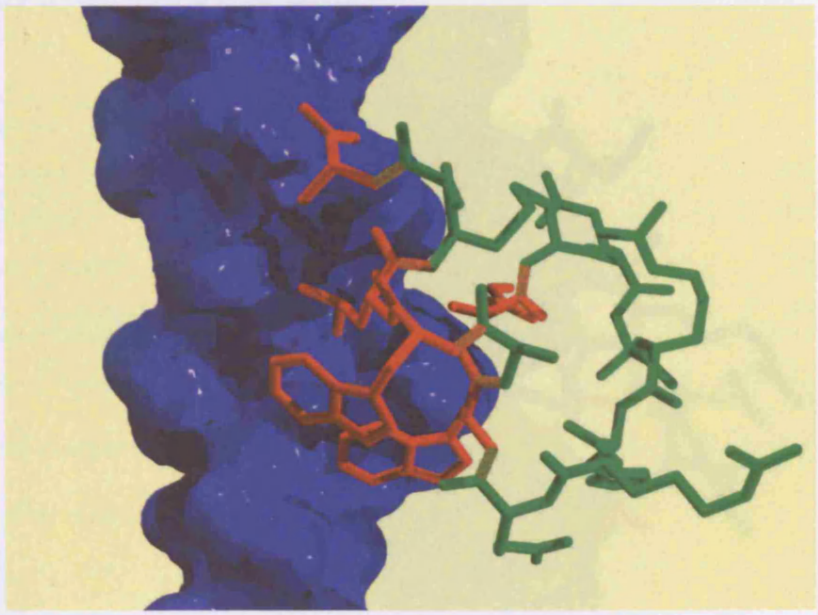

Figure 4.4: Structure of the hydrophobic pocket region in the complex formed between IQN17 and D10-p1. The residues of D10-p1 that make intimate contact with the full hydrophobic pocket are shown in red, illustrating where the missing pocket residues would be found.

modes of the D-peptide and C-peptide to the hydrophobic pocket.

Clearly, it would be desirable to study the trimeric core bound to three D-peptide or C-peptide inhibitors but valuable insights can still be gained by studying the interactions present in a monomer-inhibitor system. Scaling the system up to the trimer would not only increase the number of polypeptide atoms three-fold, it would also necessitate the use of significantly larger solvent boxes. Since a large percentage of the simulation time is already spent calculating trajectories of solvent molecules, the computational cost of such calculations is prohibitive here.

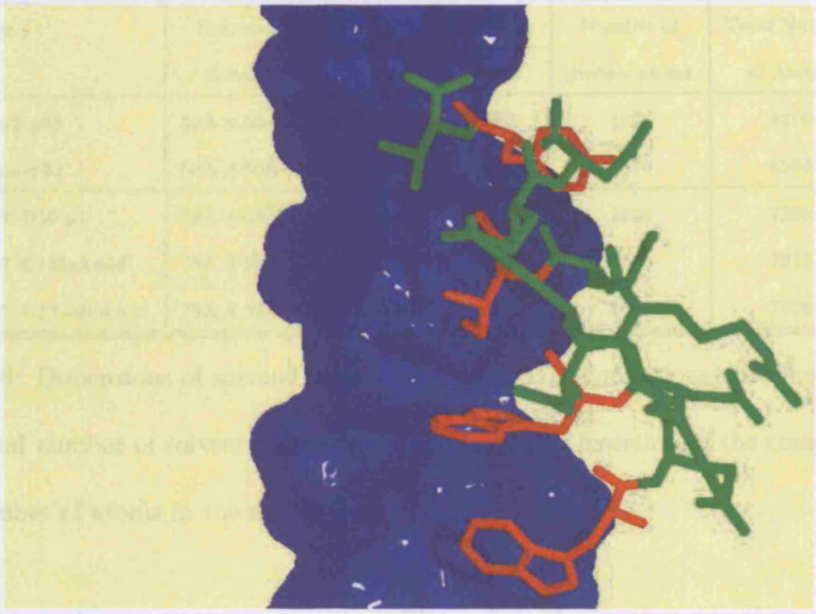

Figure 4.5: Structure of the hydrophobic pocket region in the complex formed between IQN17 and C12lm. The residues of C12lm that make intimate contact with the full hydrophobic pocket are shown in red.

4.3 Molecular Dynamics Simulations

4.3.1 Preparing the Initial System

The first stage in preparing the molecular system for the MD simulation was to create a minimised box of TIP3P[81] water molecules into which the protein–peptide system could be solvated. The size of the water box was chosen to ensure that the whole complex was well solvated but care was taken to use the smallest number of water molecules possible, in order to limit the computational cost. The complex was solvated in the water box and any water molecules that overlapped

System	Dimensions of Solvent Box	Solvent Molecules		Number of protein atoms	Total Number of Atoms
		Before	After		
xMdm2-p53	50Å x 50Å x 50Å	2600	2251	1626	8379
hMdm2-p53	50Å x 50Å x 50Å	2600	2238	1670	8384
IQN17-D10-p1	70Å x 33Å x 33Å	2400	2083	1050	7299
IQN17-C12linkmid	75Å x 33Å x 33Å	2600	2293	1040	7919
IQN17-C12unlinkmid	75Å x 33Å x 33Å	2600	2287	1045	7906

Table 4.6: Dimensions of solvent boxes used in the MD simulations described in this work, total number of solvent molecules before and after insertion of the complex and total number of atoms in the simulations.

with polypeptide atoms were removed. The solvated complex was then partially minimised (to a root mean square (RMS) gradient of 1.0 kcal mol^{-1}/Å) using a conjugate gradient minimisation algorithm within the TINKER package, in order to relieve any unfavourable atomic overlaps within the system that were still present. The dimensions of the solvent box used, as well as the number of atoms before and after insertion of the complex, are given in Table 4.6.

4.3.2 Simulation Conditions

Each of the MD simulations was carried out using a modified Beeman algorithm[217] with time steps of 2.0 fs. Bond lengths were constrained to ideal values by default using RATTLE[229] for all bonds involving hydrogen atoms. A cutoff with a polynomial switching function was used for van der Waals interactions, gradually reducing the interactions to zero between the inner and outer cutoffs: 7.2 Å and 8.0 Å, respectively. Periodic boundary conditions (PBC)[1] were used

Molecular System	Data-gathering Phase / ps
xMdm2-p53	400
hMdm2-p53	400
IQN17/D10-p1	600
IQN17/C12linkmid	400
IQN17/C12unlinkmid	350

Table 4.7: Duration / ps of each of the data-gathering phases for the different molecular systems studied in this work.

throughout the simulation. Particle mesh Ewald summation[223] was used for efficient treatment of long-range electrostatic interactions. Constant pressure and temperature of 1.0 atm and 300 K, respectively, were maintained by coupling to external baths using the method of Berendsen *et al.*[230] Both the pressure and temperature coupling parameters were 0.2 ps. The system was gradually heated to 300 K over 15 ps, allowing 5 ps per 100 K temperature increase. The system was then allowed to equilibrate at 300 K for a further 25 ps before the data-gathering phase of the simulation commenced. During the data-gathering phase the atomic structure of the complex was saved after every 1.0 ps of simulation time. The length of the data-gathering phase for each molecular system is given in Table 4.7.

4.3.3 Analysing the Trajectory

The code MDANAL was developed in this laboratory to analyse the MD trajectory. The code takes each of the saved complex structures from the data-gathering phase of the trajectory, removes the explicit TIP3P[81] water molecules

and calculates the molecular mechanical (MM) and continuum solvation (GBSA) energy for each of the complex, protein and peptide systems in the conformation found in that particular snapshot. This is repeated for all of the snapshots and average MM–GBSA energies and their components for each species are determined for the data-gathering phase of the simulation. These absolute energy values are then used to determine the binding energy, $\Delta G_{\text{subtotal}}$. The data are also analysed to yield standard deviations for the energy components in order to estimate the reliability of the data and to give insights into the stability of the MD simulation.

The program MDANAL also systematically mutates each of the peptide residues to alanine, except for those residues for which mutation is not desirable or appropriate. The sidechain atoms of the residue to be mutated are removed, with the exception of C_γ, C_β and its hydrogen atoms. C_γ is then changed to a third β-hydrogen and all of the remaining atoms of the residue are changed into their alanine equivalent. In general, this leaves the C_β–H bond distance too long, which has an effect when calculating the absolute energy and components for the complex or peptide system in question. However, since the effect on the binding energy due to changes in internal energy are defined as zero if the complex, protein and peptide structures are taken from the same snapshot of the MD trajectory, as they are in this work, the only effect observed is in the non-bonded van der Waals and electrostatic contributions to the binding energy. The difference in energy for the peptide is balanced by a corresponding difference in the complex energy and it is therefore reasonable to neglect the effect on the binding energy caused by this single bond length change. This also makes the reasonable

assumption that the continuum solvation energy is not significantly affected by this minor bond length difference.

4.4 Estimation of Entropy

4.4.1 Minimisation of the System

In order to perform the normal mode analysis of the complex, protein and peptide structures for each molecular system, it is necessary to perform an energy minimisation. It is desirable to find minima that are as close as possible to the observed X-ray structure. This can be achieved by performing the minimisation using a distance-dependent dielectric constant to mimic the effect of solvent screening.[3,174] In this work, this effect is achieved by minimising the required structure with the continuum solvation energy term activated. For comparison purposes, the minimisations are also performed without the solvent term. The root mean square (RMS) deviations in the positions of the α-carbons relative to the X-ray structure for minimisations performed both with and without the continuum solvation term applied are given in Table 4.8. These were calculated using the code CALPHA, developed in this laboratory to identify and list the α-carbons, and the SUPERPOSE program in TINKER, which superimpose two molecular structures and calculates the RMSD in the positions of the specified atoms for the best fit. This method is also used later for analysing the stability of the MD trajectories.

The calculations are expensive and it is important to ensure that there are

	Complex		Protein		Peptide	
System	No Solvent	Solvent	No Solvent	Solvent	No Solvent	Solvent
xMdm2-p53	1.79	1.73	2.33	1.75	3.83	0.57
hMdm2-p53	0.93	0.50	1.44	0.73	3.87	0.86
IQN17-D10-p1	3.61	1.58	4.46	1.39	2.16	0.61
IQN17-C12lm	3.38	1.56	4.46	1.39	1.11	0.55
IQN17-C12unlm	4.03	1.65	4.46	1.39	1.44	0.66

Table 4.8: C_α RMSD / Å, relative to the X-ray crystal structure, for each of the complex, protein and peptide of each molecular system following minimisation. The minimisations were carried out both without and with the continuum solvation energy term activated to mimic the effect of solvent screening.

six zero frequencies and no imaginary frequencies as problems arise in the calculation of the vibrational partition function and the contribution to the entropy from each mode. Since there are already significant approximations inherent in the method used here in estimating the entropy, the minimisation and normal mode analysis are performed once for IQN17 and the same data used for its complex with D10-p1, C12linkmid and C12unlinkmid. As can be seen, the structures minimised with the continuum solvation term applied are significantly closer to the X-ray configurations than those obtained without solvation. This is particularly pronounced for those systems that are not electronically neutral.

4.4.2 Entropy Codes: VIBENT and ROTRANSENT

The program VIBENT was developed in this laboratory to calculate the vibrational component of the entropy of a molecular system. Existing algorithms

within the TINKER[211] distribution are used to construct and diagonalise the Hessian matrix. The vibrational contribution to the entropy is determined using the vibrational frequencies, given by the non-zero eigenvalues of the Hessian matrix, as in Equation 3.34.

The program ROTRANSENT was developed in this laboratory to calculate the translational and rotational components of the entropy of a molecular system. Again, existing algorithms within TINKER[211] are used to determine the total mass of the molecule and the moments of inertia about its principal axes. These are used to determine the translational and rotational partition functions of the molecule, as shown in Equations 3.27 and 3.31. These are then used to determine the contributions to the entropy due to translation and rotation using Equations 3.29 and 3.33, respectively. The FORTRAN codes for ROTRANSENT and VIBENT are given in Appendix B.

4.4.3 Testing the Codes: Propane

The codes developed to determine the entropy of the polypeptides were tested on the propane molecule. The translational and rotational components of the entropy only depend on the total mass of the molecule and its moments of inertia about the principal axes, and are calculated here using the code ROTRANSENT. The most difficult contribution to evaluate is the vibrational component of the entropy, as all of the vibrational frequencies must be known. Table 4.9 gives the vibrational frequencies in cm^{-1} as assigned by Herzberg[232] based on the infrared spectrum of Wu and Barker.[235] Only twenty-five frequencies are given by Herzberg[232] and the two lowest frequency vibrations, the internal tor-

Herzberg[232]	Pitzer[233]	This Work			
ν_i / cm^{-1}	ν_i / cm^{-1}	ν_i / cm^{-1}	Contribution / kcal mol^{-1}		
			180 K	231 K	300 K
2980	2980	2969			
2968	2970	2968			
2968	2968	2967			
2967	2968	2966			
2967	2966	2944			
2960	2960	2896			
2903	2942	2875			
2885	2914	2870			
1470	1470	1472		0.0005	0.0041
1468	1468	1460		0.0005	0.0043
1465	1460	1459		0.0005	0.0044
1451	1450	1453		0.0005	0.0045
1451	1450	1450		0.0006	0.0045
1375	1375	1439		0.0006	0.0047
1370	1370	1395	0.0001	0.0007	0.0057
1278	1338	1355	0.0001	0.0009	0.0067
1179	1278	1220	0.0002	0.0020	0.0118
1152	1179	1066	0.0007	0.0046	0.0220
1152	1155	1061	0.0007	0.0047	0.0225
1053	1053	1037	0.0008	0.0054	0.0248
922	940	961	0.0014	0.0081	0.0336
870	922	938	0.0017	0.0091	0.0368
748	868	851	0.0031	0.0145	0.0520
375	748	803	0.0043	0.0186	0.0627
333	375	379	0.0726	0.1581	0.3152
	283	251	0.1635	0.2984	0.5210
	202	208	0.2141	0.3706	0.6212

Table 4.9: Vibrational frequencies / cm^{-1} of propane as assigned by Herzberg[232] and Pitzer,[233] and as determined by the code VIBENT in this work. The contribution of each frequency to the entropy, TS_{vib} / kcal mol^{-1} is also given at various temperatures.

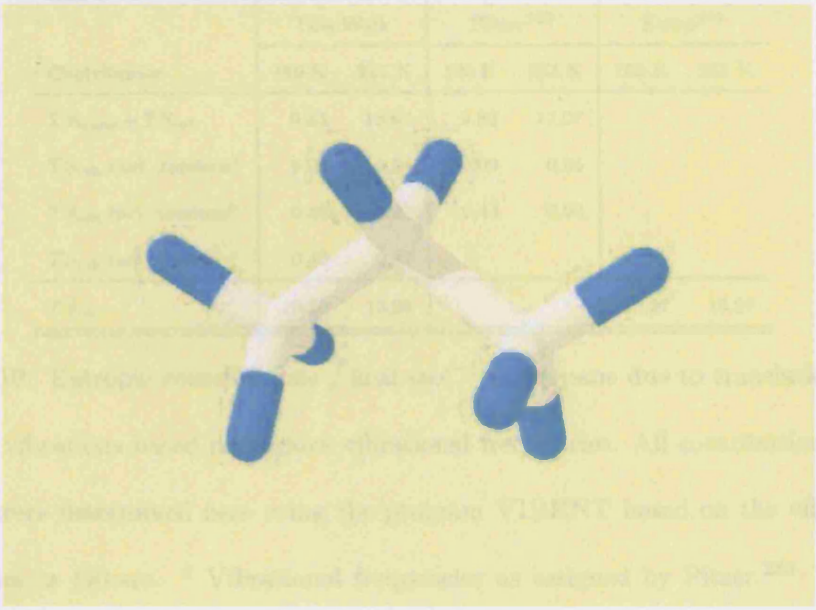

Figure 4.6: A stick representation of the three-dimensional structure of the propane molecule. Carbon atoms are shown in white and hydrogen atoms in blue. The two torsional modes referred to in the text involve rotations of the CH_3 groups about the C–C bonds. In one mode, the CH_3 groups both rotate in the same direction and in the other mode they rotate in opposite directions.

sional motions of the molecule, are not included. Pitzer[233] gives largely similar vibrational frequencies and uses the experimental entropy measurements of Kemp and Egan[234] to derive the torsional frequencies at 202 and 283 cm^{-1}. The structure of propane and a description of the torsional motions are given in Figure 4.6. The contribution to the vibrational entropy due to each normal mode are given at 180 K, 231 K and 300 K, as determined by the code VIBENT.

Table 4.10 gives the contributions to the total entropy of propane. The code

	This Work		Pitzer[233]		Kemp[234]	
Contribution	180 K	231 K	180 K	231 K	180 K	231 K
$TS_{trans} + TS_{rot}$	9.83	13.07	9.82	13.07		
TS_{vib} excl. torsions[a]	0.09	0.24	0.09	0.24		
TS_{vib} incl. torsions[b]	0.45	0.87	0.44	0.90		
TS_{vib} incl. torsions[c]	0.46	0.87				
TS_{tot}	10.30	13.98			10.27	13.97

Table 4.10: Entropic contributions / kcal mol^{-1} of propane due to translation, rotation and vibrations based on various vibrational frequencies. All contributions to the entropy were determined here using the program VIBENT based on the vibrational frequencies as follows. [a] Vibrational frequencies as assigned by Pitzer.[233] The two lowest frequency torsional vibrations are not included. [b] Vibrational frequencies as assigned by Pitzer[233] but including the derived torsional frequencies at 202 and 283 cm^{-1}. [c] Vibrational frequencies determined in this work by the program VIBENT.

VIBENT was tested using the vibrational frequencies given by Pitzer[233] to calculate the vibrational entropy. Excellent agreement is seen between the two sets of values, as would be expected. The code VIBENT was also used to determine the vibrational frequencies and excellent agreement is seen with experiment for the total entropy of propane. Pitzer[233] derived the torsional frequencies from the entropy difference between the experimental value[234] and his total entropy excluding internal rotations. The contribution to the vibrational entropy, TS_{vib} due to the two torsional modes was 0.36 kcal mol^{-1} at 180 K and 0.67 kcal mol^{-1} at 231 K. In this work, the contribution due to the torsional modes was found to be 0.38 kcal mol^{-1} at 180 K and 0.67 kcal mol^{-1} at 231 K. This shows excel-

lent agreement for the determination of the vibrational frequencies of the vital torsional modes and their contributions to the vibrational entropy.

Part III

Results and Discussion

Chapter 5

Mdm2–p53

As discussed in Chapter 1, the interactions between the oncoprotein Mdm2 and the tumour suppressor peptide p53 are of significant interest in cancer research. The interactions between *X. Laevis* and human Mdm2 (referred to as xMdm2 and hMdm2, respectively) have been investigated in this work and the results are given below.

5.1 *X. Laevis* Mdm2–p53 Interaction

5.1.1 Binding Energy

Figure 5.1 shows a plot of the root mean square deviation (RMSD) in the positions of the α-carbons in the xMdm2–p53 complex, relative to the X-ray crystal structure, over the course of the data-gathering phase of the MD simulation. The observed RMS deviations do not fluctuate greatly, suggesting that the MD simulation is stable over the timescale of interest here. The average value for C_α

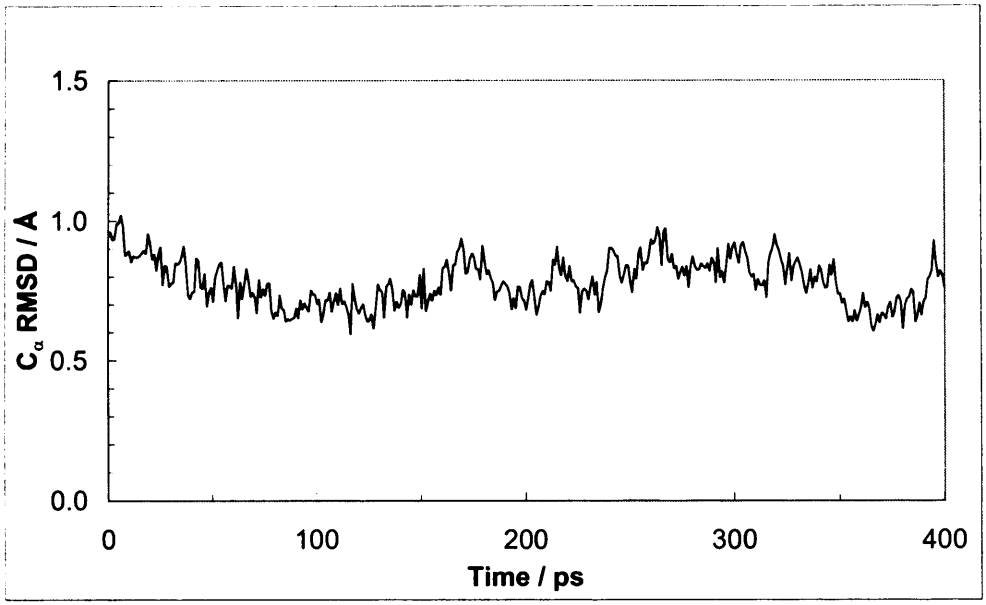

Figure 5.1: Root mean square deviation (RMSD) of the α-carbons in the xMdm2–p53 complex, relative to the X-ray crystal structure, during the data-gathering phase of the molecular dynamics trajectory.

RMSD is 0.78 Å and the standard deviation is 0.1 Å.

The absolute energy values and components for each of the complex, protein and peptide systems are given in Table 5.1. The term G_{subtotal} is simply the sum of E_{gas} and $G_{\text{solvation}}$ and is used throughout this work as the main criterion for calculating changes in binding energy. This relies on the reasonable assumption that the entropic contribution to the binding energy is the same for mutated and wild type (unmutated) peptides. The entropy and its components are given in Table 5.2 for complex, protein and peptide structures minimised both with

	Complex			Protein			Peptide		
Contribution	Mean	S.D.	M/K	Mean	S.D.	M/K	Mean	S.D.	M/K
$E_{\text{electrostatic}}$	−2815.9	43.8	−2951.6	−2512.5	33.3	−2584.8	−144.6	9.8	−265.5
E_{vdW}	−400.3	10.9	−399.0	−332.9	10.1	−331.0	−14.5	3.0	−12.8
E_{internal}	1920.5	31.8	1721.3	1708.6	29.6	1502.0	211.9	10.4	219.3
E_{gas}	−1295.7	46.8	−1629.3	−1136.8	42.5	−1413.8	52.8	14.1	−59.0
G_{polar}	−1468.8	44.3	−1653.6	−1307.8	30.8	−1509.3	−351.3	9.2	−275.3
$G_{\text{non-polar}}$	46.1	0.8	35.5	43.1	0.7	32.7	11.7	0.4	10.2
G_{subtotal}	−2718.3	35.6	−3247.3	−2401.5	33.7	−2890.4	−286.8	10.1	−324.1

Table 5.1: Absolute energy and components / kcal mol^{-1} for the xMdm2–p53 complex. Structures have been minimised with and without the application of the continuum solvation term. The columns marked M/K refer to the work of Massova and Kollman.[3]

and without the continuum solvation term applied. The binding free energy, $\Delta G_{\text{binding}}$, and its components are given in Table 5.3. The given entropic contribution to the binding free energy, $-T\Delta S$, is for structures minimised with the implicit solvent term applied. Where possible, the corresponding values from the work of Massova and Kollman[3] are given for comparison.

There are differences between the absolute energy values determined here for the complex, protein and peptide systems and those given by Massova and Kollman.[3] However, it is interesting to note that despite these differences, the calculated binding energy is very similar for the two studies. In particular, there are differences between the absolute internal energy values, E_{internal}, for the two studies but, since the complex, protein and peptide structures come from the same snapshot of the MD trajectory, $\Delta E_{\text{internal}}$ is zero by definition. The standard deviations in the absolute energy components are comparable for the two

	Complex			Protein			Peptide		
	No Solvent	Solvent	M/K	No Solvent	Solvent	M/K	No Solvent	Solvent	M/K
TS_{vib}	1066.9	1142.1		951.0	1013.0		122.3	134.6	
TS_{rot}	15.6	15.6		15.5	15.5		12.4	12.4	
TS_{trans}	16.2	16.2		16.1	16.1		14.3	14.3	
TS_{tot}	1098.7	1173.9	1198.5	982.5	1044.5	1041.9	148.9	161.3	185.0

Table 5.2: Entropy and components / kcal mol^{-1} for the xMdm2–p53 complex. The columns marked M/K refer to the work of Massova and Kollman.[3]

Contribution	Mean	S.D.	M/K
$\Delta E_{\text{electrostatic}}$	−158.7	25.9	−101.3
ΔE_{vdW}	−52.9	3.6	−55.2
ΔE_{gas}	−211.6	27.3	−155.6
ΔG_{polar}	190.2	26.1	131.0
$\Delta E_{\text{electrostatic}} + \Delta G_{polar}$	31.5	3.5	29.7
$\Delta G_{\text{non-polar}}$	−8.6	0.4	−7.8
$\Delta G_{\text{subtotal}}$	−30.0	3.5	−32.9
$-T\Delta S$ (solvent)	31.9	n/a	28.4
$\Delta G_{\text{binding}}$	1.9	n/a	−4.5

Table 5.3: Binding free energy and components / kcal mol^{-1} for the xMdm2–p53 complex. The column marked M/K refers to the work of Massova and Kollman.[3]

studies.

There are also differences between the absolute values of $E_{\text{electrostatic}}$ and G_{polar} for the two studies. In this work, the $E_{\text{electrostatic}}$ values are less negative for each of the complex, protein and peptide systems. This has the effect that for E_{gas}, the value determined here is of opposite sign to the literature value, although the

zero in energy is effectively arbitrary and the discrepancy in sign has no particular physical significance. It is worth noting that the molecular systems studied here and by Massova and Kollman[3] differ in the number of amino acid residues in the p53-derived peptide. These differences have been described in Chapter 4 and are illustrated in Table 4.2. There is a corresponding decrease in the magnitude of the G_{polar} values for the protein and peptide systems in this study compared to the work of Massova and Kollman.[3] However, the value of G_{polar} for the p53-derived peptide is determined to be more negative than that reported by Massova and Kollman.[3] This may be partly due to the fact that the N-terminal residue of the p53-derived peptide used here, Glu-1 (Glu-17 as labelled by Massova and Kollman[3]), is a charged residue that will be highly exposed to the continuum solvent giving a large G_{polar}. This compares with the polar but uncharged N-terminal glutamine of the 12-residue p53-derived peptide used by Massova and Kollman.[3] Also, the two studies use different methods to evaluate the polar contribution to the solvation free energy. This study uses the Generalized Born method while Massova and Kollman[3] use the Poisson-Boltzmann method.

Massova and Kollman[3] estimated the entropy of each of the complex, protein and peptide systems using normal mode analysis. The structures were minimised using a distance-dependent dielectric constant to mimic the effect of solvent screening and obtain minimum energy structures similar to those seen in the X-ray crystallographic structure. In this work this was achieved by minimising the structures both with and without the continuum solvation energy term applied. When minimised with the application of the solvent term, the structure is found to deviate from the crystal structure by significantly less, as shown previ-

ously in Table 4.8. The differences in C_α RMSD for the unsolvated and solvated structures are more pronounced for the charged protein and peptide systems, whereas the difference is small for the electronically neutral complex. The absolute entropy values for the solvent-minimised structure show good agreement with the values given by Massova and Kollman.[3]

The entropic contribution to the binding free energy, $-T\Delta S$, is 32.8 kcal mol^{-1} for the unsolvated structure and 31.9 kcal mol^{-1} for the solvated structure at 300K. This compares with a value of 28–36 kcal mol^{-1} given by Massova and Kollman.[3] This value depended on how well the structure was minimised before the normal mode analysis was performed. In their determination of the binding free energy, $\Delta G_{\text{binding}}$, they used a value of 28.4 kcal mol^{-1}.

Despite the differences in the absolute energy values for the complex, protein and peptide structures, the calculated binding energy, excluding entropic effects, $\Delta G_{\text{subtotal}}$, is very similar for the two studies considering the differences in the p53-derived peptides used in each case. The changes in electrostatic energy and polar solvation energy upon binding, $\Delta E_{\text{electrostatic}}$ and ΔG_{polar}, are both found to be larger in magnitude here than the values determined by Massova and Kollman.[3] These two contributions to the binding energy have opposing effects, with $\Delta E_{\text{electrostatic}}$ favouring binding and ΔG_{polar} opposing complex formation as polar residues are shielded from the continuum solvent in the protein–peptide interface. Both of these energy contributions fluctuate considerably over the timescale of the MD simulation, as indicated by their large standard deviations. However, these fluctuations tend to cancel each other out and the sum of the two terms remains reasonably constant. This is reflected by the much smaller standard de-

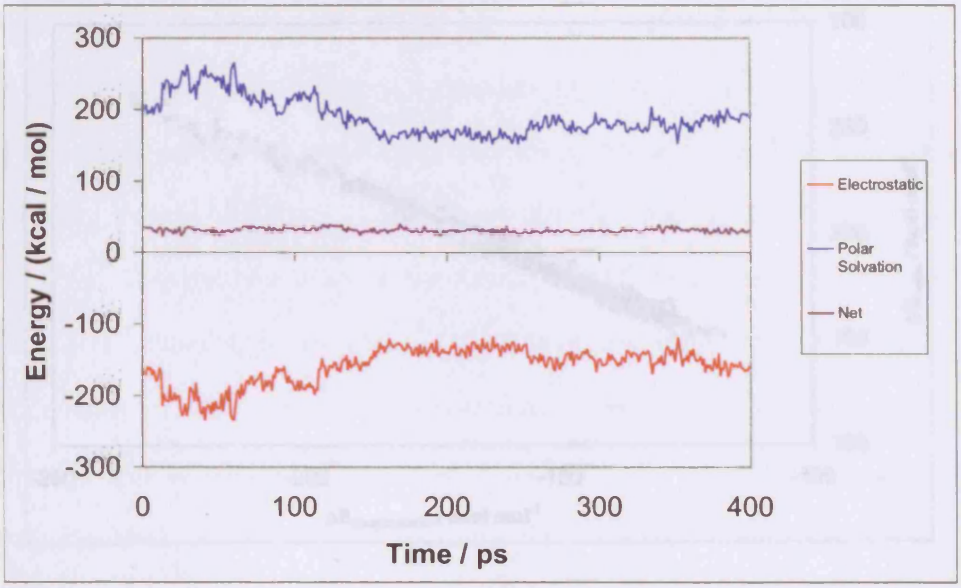

Figure 5.2: Contributions to the binding energy of the xMdm2–p53 complex due to electrostatic energy, polar solvation energy and the net effect of these two terms during the data-gathering phase of the molecular dynamics trajectory.

viation for the net term and is also illustrated in Figure 5.2. This relationship is further illustrated in Figure 5.3, a plot of $E_{\text{electrostatic}}$ against G_{polar}, which shows the strong correlation between the two contributions. The calculated correlation coefficient is -0.99, confirming the linear relationship.

The difference between $\Delta G_{\text{subtotal}}$ for this work and that of Massova and Kollman[3] is 2.9 kcal mol^{-1}. This compares with a difference of 2.3 kcal mol^{-1} in ΔE_{vdW} for the two studies. In this work there is one less residue in contact with the binding site and the corresponding reduction in the number of hydrophobic

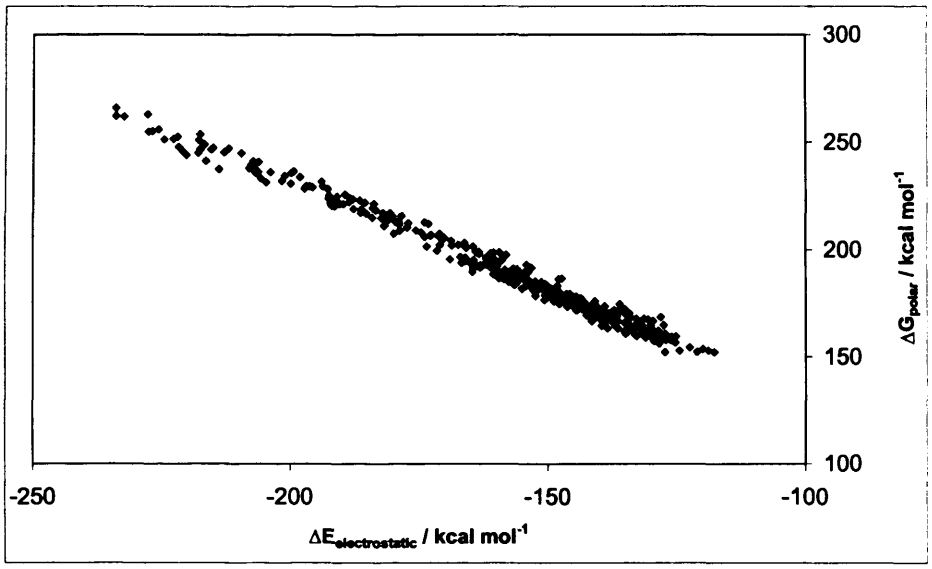

Figure 5.3: Plot of $\Delta E_{\text{electrostatic}}$ against ΔG_{polar} for the xMdm2–p53 complex. The correlation coefficient is -0.99.

van der Waals interactions leads to this difference in ΔE_{vdW}.

The calculated value of 1.9 kcal mol^{-1} for the free energy of binding, $\Delta G_{\text{binding}}$, compares to the value of -4.5 kcal mol^{-1} determined by Massova and Kollman.[3] The experimentally determined binding free energy of -6.6 to -7.8 kcal mol^{-1} is based on the reported IC_{50} value of 2–14 μM.[6] Although the small, positive value for $\Delta G_{\text{binding}}$ is of some concern, it is not overly surprising given the approximations inherent in the determination of the entropic contribution. There is good agreement between this work and the previous study[3] for $\Delta G_{\text{subtotal}}$, which will be used as the main criterion for determination of changes in binding energy upon mutation of the peptide residues to alanine.

5.1.2 Computational Alanine Scanning

Each of the residues in the p53-derived peptide, apart from Pro-11 (Pro-27), has been systematically mutated to alanine and the resulting changes in the mean binding energy and its components are given in Table 5.4. The changes in binding energy, $\Delta\Delta G_{\text{subtotal}}$, are shown for the mutation of each residue in Figure 5.4. This method relies on the assumption that only local conformational change result from each mutation. This assumption was examined by Massova and Kollman[3] by carrying out a separate simulation with the Trp23Ala mutation in the p53-derived peptide. They found that the binding energies, $\Delta G_{\text{subtotal}}$, were nearly identical for the mutated and modified wild-type trajectories, justifying the assumption. Clearly, this assumption is not reasonable for the proline to alanine mutation. A negative value of $\Delta\Delta G_{\text{subtotal}}$ indicates a mutation leading to less favourable binding of the p53-derived peptide to Mdm2, whereas a positive $\Delta\Delta G_{\text{subtotal}}$ indicates a preference for alanine at that position.

Amino acid replacement experiments[6] have identified the residues Phe-3, Trp-7 and Leu-10 (Phe-19, Trp-23 and Leu-26) as key to the binding between p53 and Mdm2. The experiments, carried out on human Mdm2 with a 12-residue p53-derived peptide, found that no other amino acid could be substituted at these positions without increasing the IC_{50} at least threefold. The residue Leu-6 (Leu-22) was also found to be highly selective, with only two other amino acids allowed without a threefold increase in IC_{50}. The calculations of Massova and Kollman[3] identified these residues as having large, negative values of $\Delta\Delta G_{\text{subtotal}}$ upon mutation to alanine, confirming them as key to the binding of p53 and Mdm2. These four residues are also found to have large, negative changes in

	Mutation								
	Glu1Ala			**Thr2Ala**			**Phe3Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−38.5	24.4	−66.3	−1.4	1.0	−1.7	−1.4	0.6	−0.8
$\Delta\Delta E_{\text{vdW}}$	−0.5	0.6	−0.7	−0.4	0.2	−0.7	−9.0	0.9	−8.8
$\Delta\Delta E_{\text{gas}}$	−38.9	24.2	−67.0	−1.8	1.2	−2.4	−10.4	1.0	−9.6
$\Delta\Delta G_{\text{polar}}$	39.7	23.7	68.2	1.3	0.8	2.2	6.8	0.6	6.2
$\Delta\Delta G_{\text{non-polar}}$	−0.1	0.1	−0.1	0.01	0.0	0.0	−0.5	0.1	−0.3
$\Delta\Delta G_{\text{subtotal}}$	0.6	1.0	1.1	−0.5	0.5	−0.2	−4.1	1.0	−3.7
	Ser4Ala			**Asp5Ala**			**Leu6Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−0.1	1.3	−0.2	−26.1	3.6	−24.2	−0.1	0.3	0.1
$\Delta\Delta E_{\text{vdW}}$	−0.4	0.2	−0.3	−0.1	0.0	−0.1	−3.1	0.6	−3.8
$\Delta\Delta E_{\text{gas}}$	−0.5	1.3	−0.5	−26.2	3.7	−24.5	−3.2	0.7	−3.7
$\Delta\Delta G_{\text{polar}}$	1.4	1.3	0.9	26.5	3.6	24.2	0.9	0.4	2.2
$\Delta\Delta G_{\text{non-polar}}$	0.0	0.0	−0.1	0.0	0.0	0.0	−0.3	0.1	−0.2
$\Delta\Delta G_{\text{subtotal}}$	0.9	0.4	0.3	0.3	0.1	−0.3	−2.6	0.6	−1.7
	Trp7Ala			**Lys8Ala**			**Leu9Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−4.6	1.4	−4.8	24.0	3.4	33.1	−0.1	0.1	−0.1
$\Delta\Delta E_{\text{vdW}}$	−12.1	1.3	−11.7	−0.3	0.1	−0.1	−0.3	0.4	−0.2
$\Delta\Delta E_{\text{gas}}$	−16.7	1.4	−16.5	23.7	3.4	33.0	−0.4	0.4	−0.4
$\Delta\Delta G_{\text{polar}}$	13.9	1.1	10.8	−23.2	3.4	−33.0	0.3	0.1	0.1
$\Delta\Delta G_{\text{non-polar}}$	−1.0	0.1	−0.6	−0.1	0.1	0.0	0.0	0.0	0.0
$\Delta\Delta G_{\text{subtotal}}$	−3.8	1.3	−6.3	0.4	0.1	0.0	−0.2	0.4	−0.3
	Leu10Ala								
Contribution	Mean	SD	M/K						
$\Delta\Delta E_{\text{electrostatic}}$	−0.3	0.2	−0.3						
$\Delta\Delta E_{\text{vdW}}$	−5.4	0.7	−3.9						
$\Delta\Delta E_{\text{gas}}$	−5.7	0.8	−4.2						
$\Delta\Delta G_{\text{polar}}$	4.0	0.5	2.7						
$\Delta\Delta G_{\text{non-polar}}$	−0.3	0.1	0.0						
$\Delta\Delta G_{\text{subtotal}}$	−2.0	0.8	−1.5						

Table 5.4: Computational alanine scanning mutagenesis energy changes / kcal mol^{-1} for the xMdm2–p53 complex. The columns marked M/K refer to the work of Massova and Kollman.[3]

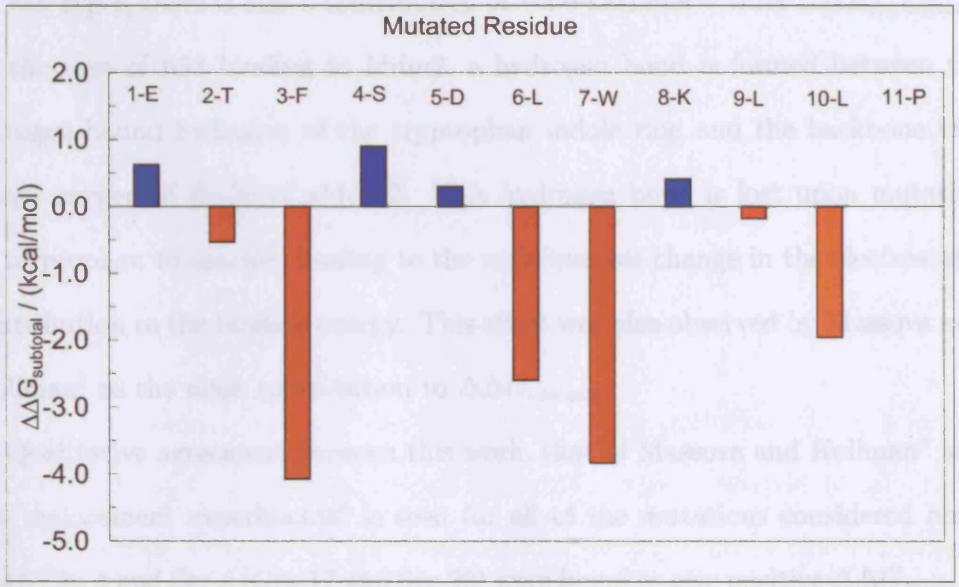

Figure 5.4: $\Delta\Delta G_{\text{subtotal}}$ / kcal mol^{-1} for the systematic mutation to alanine of each p53-derived peptide residue in the xMdm2–p53 complex. Residues marked in blue indicate favourable mutations and residues marked in red indicate unfavourable mutations.

binding energy in the present work, as shown in Table 5.4 and Figure 5.4.

The bulky sidechains of each of these four residues are involved in hydrophobic van der Waals interactions with residues in the binding site of Mdm2 that are lost when the residue is mutated to alanine. This results in a large, negative $\Delta\Delta E_{\text{vdW}}$ and opposes the mutation to alanine. There is also a gain in ΔG_{polar} upon mutation, favouring the smaller alanine residue. For Phe-3, Leu-6 and Leu-10 (Phe-19, Leu-22 and Leu-26), the loss of van der Waals interactions upon mutation is not balanced by the gain in polar solvation energy and a large, negative $\Delta\Delta G_{\text{subtotal}}$

results.

For Trp-7, there is also a contribution of -4.6 kcal mol^{-1} from $\Delta\Delta E_{\text{electrostatic}}$. In the case of p53 binding to Mdm2, a hydrogen bond is formed between the nitrogen-bound hydrogen of the tryptophan indole ring and the backbone carbonyl oxygen of Ile-50 of xMdm2. This hydrogen bond is lost upon mutation of tryptophan to alanine, leading to the unfavourable change in the electrostatic contribution to the binding energy. This effect was also observed by Massova and Kollman[3] as the main contribution to $\Delta\Delta G_{\text{subtotal}}$.

Qualitative agreement between this work, that of Massova and Kollman[3] and the replacement experiments[6] is seen for all of the mutations considered here. Both Glu-1 and Ser-4 (Glu-17 and Ser-20) were found to give positive $\Delta\Delta G_{\text{subtotal}}$, indicating a slight preference for alanine at those positions. This is caused by replacing a polar residue with a hydrophobic alanine and, in fact, the best binding sequence determined by Böttger *et al.*[6] had a more hydrophobic residue (proline and methionine, respectively) at each of these positions. Small, almost negligible, positive $\Delta\Delta G_{\text{subtotal}}$ were observed for the mutations of Asp-5 and Lys-8 (Asp-21 and Lys-24). The mutations of Thr-2 and Leu-9 (Thr-18 and Leu-25) gave rise to small, almost negligible negative $\Delta\Delta G_{\text{subtotal}}$, as also observed by Massova and Kollman.[3]

In general, the energy terms can be considered in the context of their physical origin. Thus, terms of an electrostatic origin, $\Delta\Delta E_{\text{electrostatic}}$ and $\Delta\Delta G_{\text{polar}}$, can be considered as a single contribution. For residues with a large, negative $\Delta\Delta G_{\text{subtotal}}$ the main contribution is due to the unfavourable $\Delta\Delta E_{\text{vdW}}$ not being balanced by the net electrostatic contribution, $\Delta\Delta E_{\text{electrostatic}} + \Delta\Delta G_{\text{polar}}$.

The effect of $\Delta\Delta G_{\text{non-polar}}$ was found to be negligible for most of the mutations considered here, although a small, unfavourable contribution was found for the Trp7Ala mutation.

5.2 Human Mdm2–p53 Interaction

5.2.1 Binding Energy

Figure 5.5 shows a plot of the root mean square deviation (RMSD) in the positions of the α-carbons in the hMdm2–p53 complex, relative to the X-ray crystal structure, over the course of the data-gathering phase of the MD simulation. The RMSD values are higher than for the xMdm2–p53 complex and some instability is seen, as also observed by Massova and Kollman.[3] The net charge of $+3e$ may contribute to this as systems with a net charge can exhibit artificial behaviour when PME is used, particularly under constant pressure conditions.[225] The average value for C_α RMSD is 1.01 Å and the standard deviation is 0.1 Å.

The absolute energy values and components for each of the complex, protein and peptide systems are given in Table 5.5. The entropy and its components are given in Table 5.6 for complex, protein and peptide structures minimised both with and without the continuum solvation term applied. The binding free energy, $\Delta G_{\text{binding}}$, and its components are given in Table 5.7. The given entropic contribution to the binding free energy, $-T\Delta S$, is for structures minimised with the implicit solvent term applied. Where possible, the corresponding values from the work of Massova and Kollman[3] are given for comparison. Once again, the

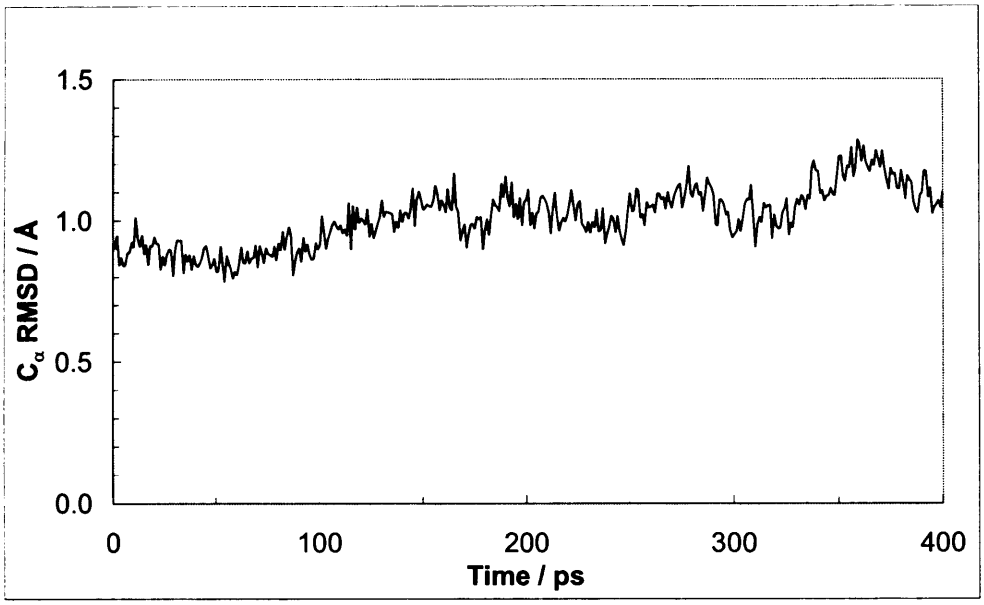

Figure 5.5: Root mean square deviation (RMSD) of the α-carbons in the hMdm2–p53 complex, relative to the X-ray crystal structure, during the data-gathering phase of the molecular dynamics trajectory.

main criterion used in determining changes in binding energy is $\Delta G_{\text{subtotal}}$.

There are differences between the p53-derived peptide used in this work and that used by Massova and Kollman.[3] These are summarised in Table 4.2. There are extra residues in the peptide used in this study and the net charge of the complex is $+3e$ compared to $+4e$ for the work of Massova and Kollman.[3] Hence, it is not surprising to see discrepancies between the two studies for the terms of electrostatic origin, $E_{\text{electrostatic}}$ and G_{polar}. The standard deviations for these terms are also larger for the human Mdm2–p53 system than for the electronically

	Complex			Protein			Peptide		
Contribution	Mean	S.D.	M/K	Mean	S.D.	M/K	Mean	S.D.	M/K
$E_{\text{electrostatic}}$	−3040.5	61.6	−2875.4	−2461.5	38.6	−2517.7	−215.9	23.4	−255.1
E_{vdW}	−393.1	12.1	−385.5	−307.5	10.8	−313.1	−11.7	3.2	−14.6
E_{internal}	1935.8	28.5	1688.1	1689.4	26.4	1466.6	246.4	10.8	221.6
E_{gas}	−1497.9	63.3	−1572.8	−1079.6	44.2	−1364.2	18.7	25.0	−48.2
G_{polar}	−1412.2	57.0	−1731.6	−1369.0	36.7	−1577.7	−441.8	22.0	−287.7
$G_{\text{non-polar}}$	45.7	0.8	35.6	43.6	0.7	32.8	13.7	0.4	10.3
G_{subtotal}	−2894.4	27.9	−3268.9	−2405.1	25.5	−2909.2	−409.5	10.2	−325.6

Table 5.5: Absolute energy and components / kcal mol^{-1} for the hMdm2–p53 complex. The columns marked M/K refer to the work of Massova and Kollman.[3]

neutral xMdm2–p53 system studied in this work.

As seen for the xMdm2–p53 system, $E_{\text{electrostatic}}$ is smaller for the peptide and protein systems compared to the work of Massova and Kollman.[3] However, the corresponding value for the complex is larger in this work than that given by Massova and Kollman.[3] This is a reflection of the fact that the p53-derived peptide has a charge of $+2e$ in this study compared to $+e$ and that there will be stronger electrostatic interactions between the protein and peptide. As also seen for the xMdm2–p53 system, G_{polar} is smaller for the complex and protein systems compared to the work of Massova and Kollman.[3] G_{polar} is larger for the peptide in this study, which is again due to the fact that the p53-derived peptide used here has a larger net charge.

As for the xMdm2–p53 system, the entropic contribution to the binding free energy has been estimated using normal mode analysis of complex, protein and peptide systems minimised both with and without the continuum solvation en-

	Complex		Protein		Peptide	
	No Solvent	Solvent	No Solvent	Solvent	No Solvent	Solvent
TS_{vib}	1074.7	1160.0	942.7	1004.6	149.6	164.1
TS_{rot}	15.6	15.6	15.4	15.5	12.7	12.9
TS_{trans}	16.2	16.2	16.1	16.1	14.4	14.4
TS_{tot}	1106.5	1191.8	974.2	1036.1	176.7	191.4

Table 5.6: Entropy and components / kcal mol^{-1} for the hMdm2–p53 complex. Structures have been minimised with and without application of the continuum solvation term.

ergy term switched on. Once again, the structure minimised with the solvent term applied deviated considerably less from the X-ray crystal structure. For the unsolvated structure, $-T\Delta S$ is 44.4 kcal mol^{-1} and for the solvated structure, $-T\Delta S$ is 35.7 kcal mol^{-1}. The difference between the entropic contributions for the two minimised structures is much more pronounced than for the xMdm2–p53 system. Once again, this could be due to the charged nature of the complex. Minimising the structure without the continuum solvent applied leads to a structure that deviates from the X-ray crystallographic structure by much more than with the solvent term applied, as shown previously in Table 4.8. The xMdm2–p53 and hMdm2–p53 systems would be expected to have roughly similar entropic contributions to the binding energy and much better agreement is seen here between the structures minimised with the continuum solvation term than for the unsolvated minimised structures.

The value determined here for the binding energy, $\Delta G_{\text{subtotal}}$, is much larger (i.e. more negative) than that determined by Massova and Kollman.[3] Once again,

Contribution	Mean	S.D.	M/K
$\Delta E_{\text{electrostatic}}$	−363.1	26.8	−102.6
ΔE_{vdW}	−73.9	4.2	−57.9
ΔE_{gas}	−437.0	26.0	−160.4
ΔG_{polar}	398.6	25.4	133.8
$\Delta E_{\text{electrostatic}} + \Delta G_{polar}$	35.6	4.52	31.2
$\Delta G_{\text{non-polar}}$	−11.6	0.4	−7.5
$\Delta G_{\text{subtotal}}$	−49.9	3.9	−34.1
$-T\Delta S$ (solvent)	35.7	n/a	
$\Delta G_{\text{binding}}$	−14.2	n/a	

Table 5.7: Binding free energy and components / kcal mol^{-1} for the hMdm2–p53 complex. The column marked M/K refers to the work of Massova and Kollman.[3]

this is likely to be due to the differences in the p53-derived peptide used in each study. As seen for the xMdm2–p53 system, the difference between $\Delta G_{\text{subtotal}}$ for the two studies, 15.8 kcal mol^{-1}, is very similar to the difference between ΔE_{vdW} for the two studies, 16.1 kcal mol^{-1}. Clearly, the larger 13-residue p53-derived peptide used in this study is involved in more van der Waals interactions with residues in the protein–peptide binding interface, leading to this stabilising effect.

As seen for the xMdm2–p53 system, $\Delta E_{\text{electrostatic}}$ and ΔG_{polar} are both larger in magnitude than the values determined by Massova and Kollman,[3] although the effect is much more pronounced here, possibly due to the presence of the extra charged residue in the p53-derived peptide used here. Despite the larger standard deviations seen for the absolute electrostatic and polar solvation terms, the standard deviations for the corresponding energy changes upon binding are comparable for the *X. Laevis* and human systems. The fluctuations tend to

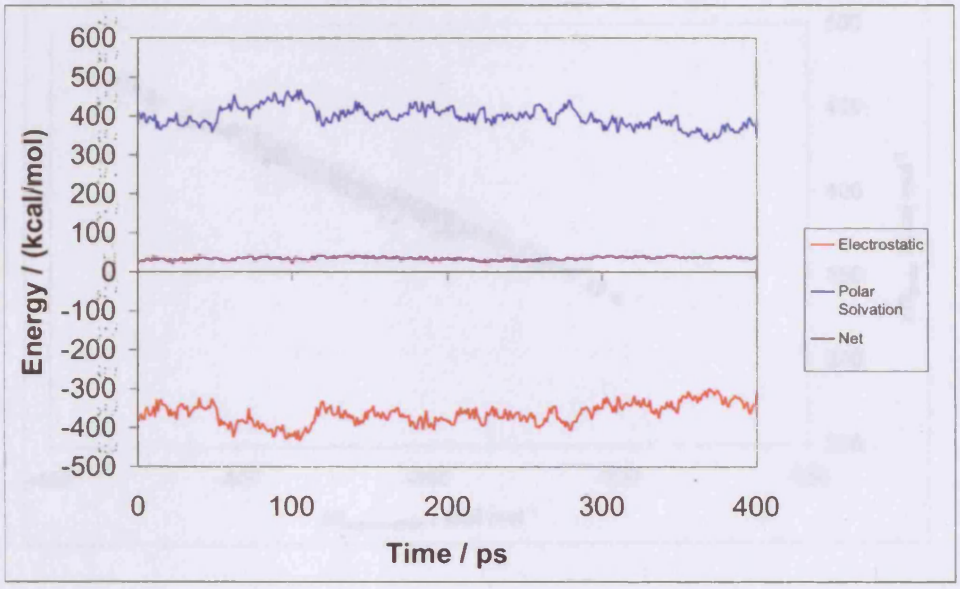

Figure 5.6: Contributions / kcal mol^{-1} to the binding energy of the hMdm2–p53 complex due to electrostatic energy, polar solvation energy and the net effect of these two terms during the data-gathering phase of the molecular dynamics trajectory.

cancel each other out, as before, and the sum of the two terms remains reasonably constant. This is supported by the relatively small standard deviation for the net term, $\Delta E_{\text{electrostatic}} + \Delta G_{\text{polar}}$, and is illustrated in Figure 5.6. This relationship is further illustrated in Figure 5.7, a plot of $E_{\text{electrostatic}}$ against G_{polar}, which shows the strong correlation between the two contributions. As for the xMdm2–p53 system, the calculated correlation coefficient is -0.99, confirming the linear relationship.

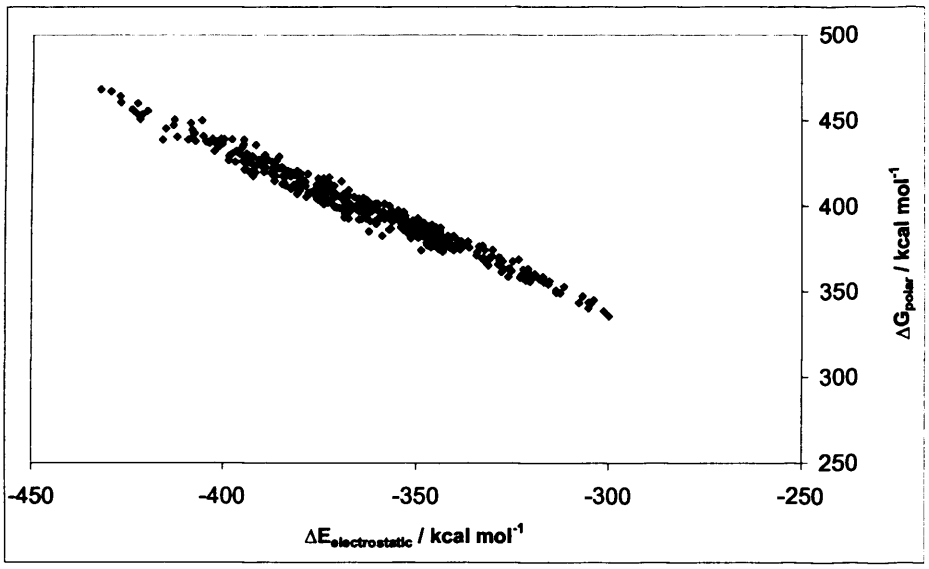

Figure 5.7: Plot of $\Delta E_{\text{electrostatic}}$ against ΔG_{polar} for the hMdm2–p53 complex. The correlation coefficient is -0.99.

5.2.2 Computational Alanine Scanning

As before, each of the residues of the p53-derived peptide, apart from Pro-11 (Pro-27), has been systematically mutated to alanine and the changes in the mean binding energy and its components are given in Table 5.8. The changes in the binding energy, $\Delta\Delta G_{\text{subtotal}}$, are shown for the mutation of each residue in Figure 5.8. As before, a negative value for $\Delta\Delta G_{\text{subtotal}}$ indicates less favourable binding for the alanine mutant, whereas a positive $\Delta\Delta G_{\text{subtotal}}$ indicates a preference for alanine at that position in the p53-derived peptide.

	Mutation								
	Glu1Ala			**Thr2Ala**			**Phe3Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−112.5	21.3	−76.3	−1.1	0.7	−1.5	−1.5	0.6	−1.2
$\Delta\Delta E_{\text{vdW}}$	−0.9	0.8	−0.9	−0.7	0.3	−0.4	−9.1	0.9	−9.2
$\Delta\Delta E_{\text{gas}}$	−113.4	21.2	−77.2	−1.8	0.8	−1.9	−10.7	0.9	−10.4
$\Delta\Delta G_{\text{polar}}$	113.1	20.5	77.2	0.5	0.7	1.8	7.4	0.7	7.9
$\Delta\Delta G_{\text{non-polar}}$	−0.1	0.1	−0.1	0.00	0.0	0.0	−0.5	0.1	−0.1
$\Delta\Delta G_{\text{subtotal}}$	−0.4	1.2	−0.1	−1.3	0.5	−0.0	−3.8	1.0	−2.7
	Ser4Ala			**Asp5Ala**			**Leu6Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−0.3	0.7	−0.2	−82.1	4.0	−68.5	−0.1	0.3	0.3
$\Delta\Delta E_{\text{vdW}}$	−0.3	0.2	−0.4	−0.1	0.0	−0.1	−4.2	0.9	−4.1
$\Delta\Delta E_{\text{gas}}$	−0.6	0.7	−0.6	−82.2	4.0	−68.6	−4.4	0.9	−3.8
$\Delta\Delta G_{\text{polar}}$	1.8	0.6	1.3	81.7	4.1	68.1	1.0	0.4	2.5
$\Delta\Delta G_{\text{non-polar}}$	0.0	0.0	0.0	0.0	0.0	0.0	−0.3	0.1	−0.2
$\Delta\Delta G_{\text{subtotal}}$	1.2	0.5	0.7	−0.5	0.2	−0.5	−3.64	0.9	−1.5
	Trp7Ala			**Lys8Ala**			**Leu9Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−5.2	1.4	−5.2	86.7	6.6	76.8	0.0	0.2	−0.1
$\Delta\Delta E_{\text{vdW}}$	−12.0	1.2	−11.0	−0.3	0.1	−0.2	−2.1	0.8	−0.6
$\Delta\Delta E_{\text{gas}}$	−17.3	1.5	−16.3	86.4	6.6	76.7	−2.1	1.0	−0.7
$\Delta\Delta G_{\text{polar}}$	14.3	1.4	11.4	−85.1	6.5	−76.4	0.6	0.3	0.7
$\Delta\Delta G_{\text{non-polar}}$	−0.9	0.1	−0.7	−0.1	0.2	−0.0	−0.2	0.1	−0.1
$\Delta\Delta G_{\text{subtotal}}$	−3.9	1.1	−5.5	1.2	0.2	0.3	−1.7	0.8	−0.0
	Leu10Ala			**Glu12Ala**			**Asn13Ala**		
Contribution	Mean	SD	M/K	Mean	SD	M/K	Mean	SD	M/K
$\Delta\Delta E_{\text{electrostatic}}$	−0.1	0.2	0.2	−158.0	20.3		−15.2	5.9	
$\Delta\Delta E_{\text{vdW}}$	−5.6	0.8	−5.1	−0.4	1.3		−1.3	1.1	
$\Delta\Delta E_{\text{gas}}$	−5.7	0.8	−4.9	−158.4	19.9		−16.6	5.7	
$\Delta\Delta G_{\text{polar}}$	3.9	0.4	3.3	156.6	19.0		13.3	3.6	
$\Delta\Delta G_{\text{non-polar}}$	−0.2	0.1	−0.1	−0.2	0.1		−0.3	0.1	
$\Delta\Delta G_{\text{subtotal}}$	−2.0	0.8	−1.7	−2.1	1.5		−3.6	2.7	

Table 5.8: Computational alanine scanning mutagenesis energy changes and components / kcal mol^{-1} for the hMdm2–p53 complex. Columns marked M/K refer to the work of Massova and Kollman.[3]

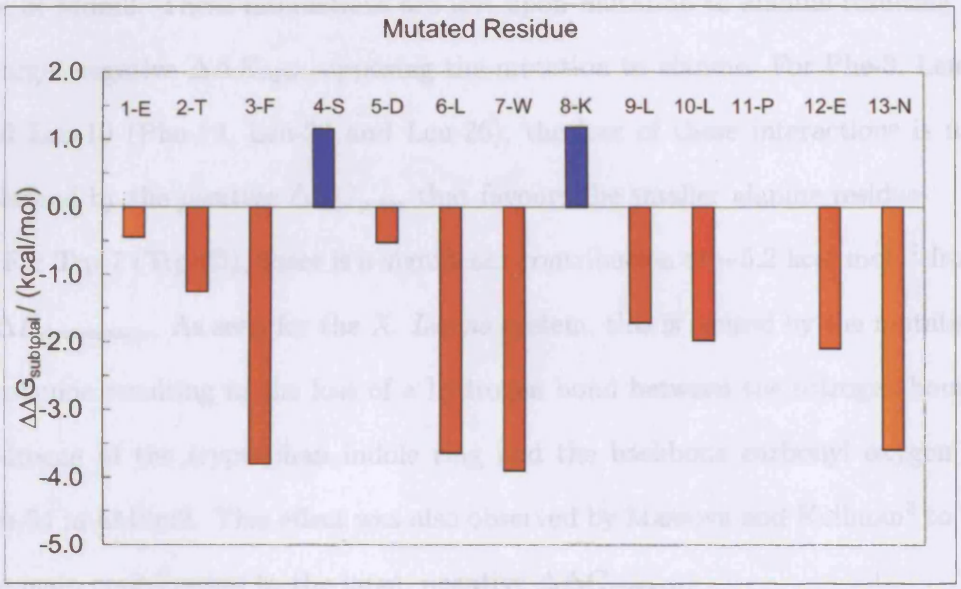

Figure 5.8: $\Delta\Delta G_{\text{subtotal}}$ / kcal mol^{-1} for the systematic mutation to alanine of each p53-derived peptide residue in the hMdm2–p53 complex. Residues marked in blue indicate favourable mutations and residues marked in red indicate unfavourable mutations.

As discussed previously, amino acid replacement experiments[6] on human Mdm2 bound to a 12-residue p53-derived peptide have identified several residues, Phe-3, Leu-6, Trp-7 and Leu-10 (Phe-19, Leu-22, Trp-23 and Leu-26), as being key to the interaction. These residues were also identified by Massova and Kollman[3] as giving large, negative values for $\Delta\Delta G_{\text{subtotal}}$ upon mutation to alanine. As for the *X. Laevis* system, the present work found large, negative $\Delta\Delta G_{\text{subtotal}}$ values for these residues, as shown in Table 5.8 and Figure 5.8.

As also discussed previously, these residues all have bulky sidechains that

participate in hydrophobic van der Waals interactions with residues in the binding site of Mdm2. These interactions are lost upon mutation to alanine resulting in a large, negative $\Delta\Delta E_{\text{vdW}}$, opposing the mutation to alanine. For Phe-3, Leu-6 and Leu-10 (Phe-19, Leu-22 and Leu-26), the loss of these interactions is not balanced by the positive $\Delta\Delta G_{\text{polar}}$ that favours the smaller alanine residue.

For Trp-7 (Trp-23), there is a significant contribution of -5.2 kcal mol^{-1} from $\Delta\Delta E_{\text{electrostatic}}$. As seen for the *X. Laevis* system, this is caused by the mutation to alanine resulting in the loss of a hydrogen bond between the nitrogen-bound hydrogen of the tryptophan indole ring and the backbone carbonyl oxygen of Leu-54 in hMdm2. This effect was also observed by Massova and Kollman[3] to be the main contribution to the large, negative $\Delta\Delta G_{\text{subtotal}}$.

Unlike for the xMdm2–p53 system, a signifcant change in binding energy is observed for the mutation of Leu-9 (Leu-25). The main contribution to $\Delta\Delta G_{\text{subtotal}}$ is $\Delta\Delta E_{\text{vdW}}$. No such loss of van der Waals interactions is observed in the *X. Laevis* system, probably due to the presence of the two extra residues, Glu-12 and Asn-13 (Glu-28 and Asn-29), not present in the system studied by Massova and Kollman.[3] A plot of $\Delta\Delta E_{\text{vdW}}$ against time for the human and *X. Laevis* systems is given in Figure 5.9. For the xMdm2–p53 system, the loss of van der Waals interactions is small and fairly constant throughout the simulation. For the hMdm2–p53 system, $\Delta\Delta E_{\text{vdW}}$ behaves similarly at the start of the simulation but then increases and fluctuates significantly as time increases. The standard deviation is twice as large for the human system. This suggests that the p53-derived peptide behaves slightly differently in the binding site of Mdm2 during the course of the MD simulation, bringing the leucine sidechain into greater contact with

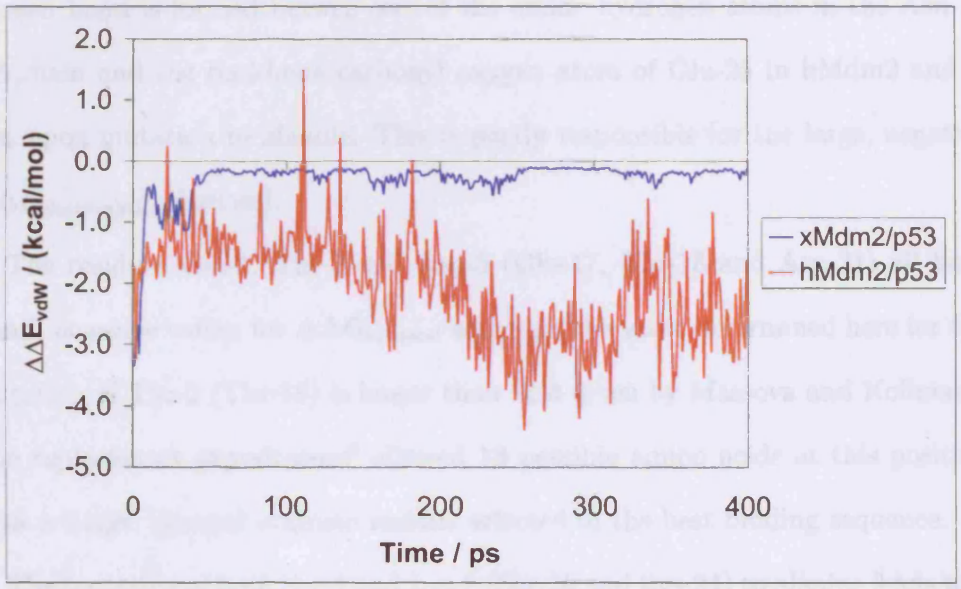

Figure 5.9: Comparison of $\Delta\Delta E_{\mathrm{vdW}}$ for the Leu9Ala mutation in each of the xMdm2–p53 and hMdm2–p53 systems over the data-gathering phase of the molecular dynamics trajectory.

the binding site.

Both of the extra residues, Glu-12 and Asn-13 (Glu-28 and Asn-29), are highly polar and give large, negative $\Delta\Delta G_{\mathrm{subtotal}}$ upon mutation to alanine. In both cases, the net electrostatic term, $\Delta\Delta E_{\mathrm{electrostatic}} + \Delta\Delta G_{\mathrm{polar}}$, is negative and opposes the mutation to alanine. These combine with small losses of van der Waals interaction energy to give the negative $\Delta\Delta G_{\mathrm{subtotal}}$. There are two hydrogen bonds present between Asn-13 (Asn-29) and Mdm2. One such bond is formed between the hydroxyl hydrogen of Thr-26 in hMdm2 and the backbone carbonyl

oxygen atom of Asn-13 (Asn-29) in the p53-derived peptide. The second hydrogen bond is formed betwen one of the amide hydrogen atoms in the Asn-13 sidechain and the backbone carbonyl oxygen atom of Glu-25 in hMdm2 and is lost upon mutation to alanine. This is partly responsible for the large, negative $\Delta\Delta E_{\text{electrostatic}}$ observed.

The residues Glu-1, Thr-2 and Asp-5 (Glu-17, Thr-18 and Asp-21) all have small, negative values for $\Delta\Delta G_{\text{subtotal}}$ although the value determined here for the mutation of Thr-2 (Thr-18) is larger than that given by Massova and Kollman.[3] The replacement experiments[6] allowed 13 possible amino acids at this position with a larger, charged arginine residue selected in the best binding sequence.

The mutation of both Ser-4 and Lys-8 (Ser-20 and Lys-24) to alanine leads to a small, positive $\Delta\Delta G_{\text{subtotal}}$, as also predicted by Massova and Kollman.[3] Neither residue has a large loss in van der Waals interaction energy upon mutation as they do not participate strongly in such interactions with residues in the binding site of Mdm2. The net positive change in the electrostatic terms, $\Delta\Delta E_{\text{electrostatic}} + \Delta\Delta G_{\text{polar}}$, leads to the slight preference for alanine at these positions. In the replacement experiments,[6] the polar serine residue is replaced by a hydrophobic methionine and the positively charged lysine residue is replaced by a negatively charged glutamic acid. Therefore, it is not surprising to see a slight preference for alanine at these positions.

Chapter 6

IQN17

As discussed in Chapter 1, IQN17 is a synthetic peptide designed to represent a potential drug target in the gp41-mediated HIV cellular fusion process. The hydrophobic pocket region is of particular interest as it is essentially conserved in the vast majority of strains of HIV-1.[21] The interactions between IQN17 and a D-peptide, D10-p1, which has been shown to inhibit the cellular fusion process have been investigated in this work and the results are given below. The interactions between IQN17 and two C-peptides, C12linkmid (C12lm) and C12unlinkmid (C12unlm), have also been investigated.

6.1 IQN17–D10-p1 Interaction

6.1.1 Binding Energy

Figure 6.1 shows a plot of the root mean square deviation (RMSD) in the positions of the α-carbons in the IQN17–D10-p1 complex, relative to the X-ray

crystal structure,[21] over the course of the data-gathering phase of the MD trajectory. Clearly, the deviations from the crystal structure are significantly larger than for either of the Mdm2–p53 systems. The net charge of $+4e$ may contribute to this as systems with a net charge can exhibit artificial behaviour when PME is used, particularly under constant pressure conditions.[225] Furthermore, the attempts to limit the system size by using the smallest possible solvent box may also have contributed to the observed instability. It is likely that more stable simulations would have been achieved using counterions to make the system electronically neutral and using a larger solvent box to ensure that the entire system was surrounded by a more significant amount of solvent at all times. The shape of the protein IQN17, a long α-helix, made it difficult to completely solvate the whole molecule without using a very large solvent box. Therefore, although the complex was smaller than the Mdm2–p53 complexes discussed previously in terms of the number of atoms, more solvent molecules would have been needed to achieve the same depth of solvent around the complex. The average value for C_α RMSD is 1.64 Å and the standard deviation is 0.4 Å.

The absolute energy values and components for each of the complex, protein and peptide systems are given in Table 6.1. The entropy and its components are given in Table 6.2 for complex, protein and peptide structures minimised with and without the continuum solvation energy term applied. The binding free energy, $\Delta G_{\text{binding}}$, and its components are given in Table 6.3. The given entropic contribution to the binding free energy, $-T\Delta S$, is for structures minimised with the implicit solvent term applied.

The complex formed between IQN17 and D10-p1 is considerably smaller than

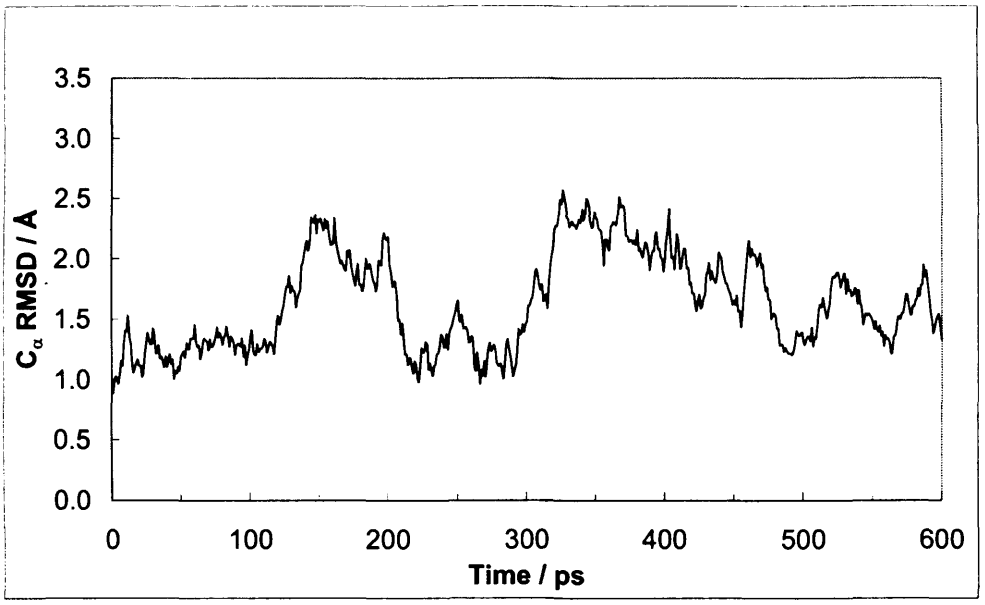

Figure 6.1: Root mean square deviation (RMSD) of the α-carbons in the IQN17–D10-p1 complex, relative to the X-ray crystal structure, during the data-gathering phase of the molecular dynamics trajectory.

the Mdm2–p53 systems considered in Chapter 5, as can be seen from a comparison of Tables 4.2 and 4.5. Unsurprisingly, the absolute energy values for the complex and protein systems are smaller than those determined for the Mdm2–p53 systems but the 16-residue D-peptide has larger absolute energy values than either the 11-residue or 13-residue p53-derived peptide. The standard deviations in the absolute energy terms are comparable for the different systems.

The entropic contribution to the binding free energy, $-T\Delta S$, is estimated to be 21.0 kcal mol^{-1} for the structures minimised without the continuum solva-

	Complex		Protein		Peptide	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.
$E_{\text{electrostatic}}$	−1750.0	48.9	−1176.5	48.0	−473.5	12.0
E_{vdW}	−179.5	7.9	−114.2	6.9	−34.5	3.9
E_{internal}	1192.3	26.1	908.4	22.8	283.8	12.7
E_{gas}	−737.1	51.2	−382.3	48.8	−224.2	16.0
G_{polar}	−1643.8	46.9	−1382.5	47.6	−379.1	10.3
$G_{\text{non-polar}}$	45.5	0.7	37.9	0.6	13.4	0.4
G_{subtotal}	−2335.41	24.8	−1726.92	21.3	−589.88	12.0

Table 6.1: Absolute energy and components / kcal mol^{-1} for the IQN17–D10-p1 complex.

	Complex		Protein		Peptide	
	No Solvent	Solvent	No Solvent	Solvent	No Solvent	Solvent
TS_{vib}	701.9	762.9	539.4	580.1	156.9	172.6
TS_{rot}	15.6	15.6	15.3	15.2	12.7	12.9
TS_{trans}	15.8	15.8	15.5	15.5	14.5	14.5
TS_{tot}	733.3	794.3	570.2	610.8	184.1	199.9

Table 6.2: Entropy and components / kcal mol^{-1} for the IQN17–D10-p1 complex. Structures have been minimised with and without application of the continuum solvation term.

tion term and 16.5 kcal mol^{-1} for the solvated minimised structures at 300 K. The experimentally determined value for the formation of the complex between IQN17 and D10-p1-2K is 9.7 kcal mol^{-1} at 298 K.[236] The peptide D10-p1-2K has the same sequence as D10-p1 but with two lysine residues added to increase solubility.[21] The agreement with experiment for the structures minimised with

Contribution	Mean	SD
$\Delta E_{\text{electrostatic}}$	−100.0	24.6
ΔE_{vdW}	−30.7	2.9
ΔE_{gas}	−130.6	23.7
ΔG_{polar}	117.8	23.1
$\Delta E_{\text{electrostatic}} + \Delta G_{\text{polar}}$	17.9	3.1
$\Delta G_{\text{non-polar}}$	−5.8	3.1
$\Delta G_{\text{subtotal}}$	−18.6	3.2
$-T\Delta S$ (solvent)	16.5	n/a
$\Delta G_{\text{binding}}$	-2.1	n/a

Table 6.3: Binding energy and components / kcal mol^{-1} for the IQN17–D10-p1 complex.

the implicit solvent present is reasonable considering the approximations inherent in the method.

As seen for the Mdm2–p53 systems, the contributions to the binding energy arising from terms of electrostatic origin, $\Delta E_{\text{electrostatic}}$ and ΔG_{polar} fluctuate considerably throughout the course of the MD trajectory, as suggested by their large standard deviations. However, as before, these fluctuations tend to cancel one another out and the net effect of the two terms remains reasonably constant. This is reflected by the smaller standard deviation for the net term and is also illustrated in Figure 6.2. This relationship is further illustrated in Figure 6.3, a plot of $E_{\text{electrostatic}}$ against G_{polar}, which shows the strong correlation between the two contributions. The calculated correlation coefficient is -0.99, again confirming the linear relationship.

The experimentally determined enthalpy change, ΔH^0, for the formation of

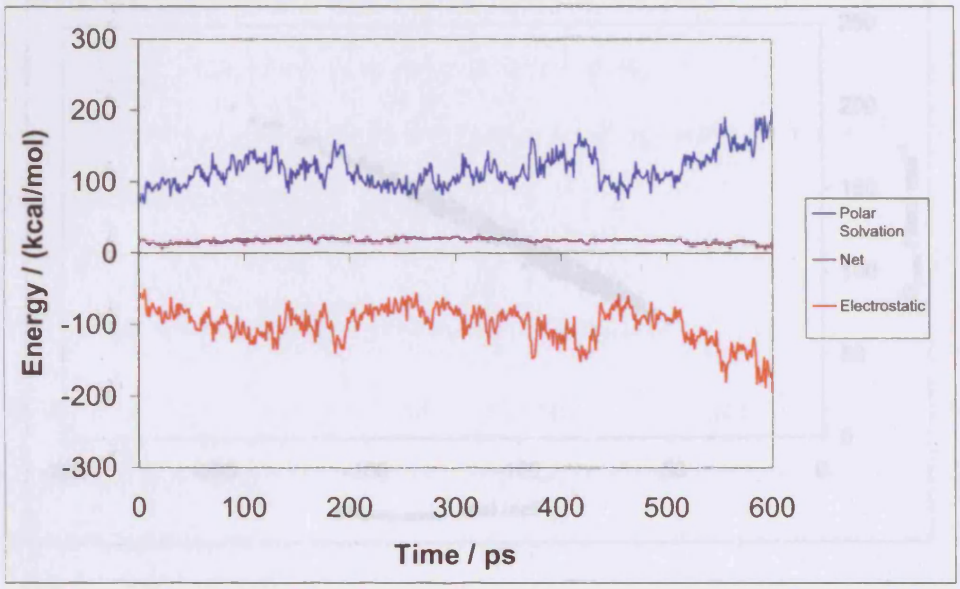

Figure 6.2: Contributions to the binding energy of the IQN17–D10-p1 complex due to electrostatic energy, polar solvation energy and the net effect of these two terms during the data-gathering phase of the molecular dynamics trajectory.

the IQN17–D10-p1-2K complex is -16.6 kcal mol^{-1} at 298 K.[236] Although this quantity does not have exactly the same physical meaning as $\Delta G_{\text{subtotal}}$, the significance is similar and good agreement is seen between this work and experiment. The calculated binding free energy, $\Delta G_{\text{binding}}$, is -2.1 kcal mol^{-1}, which is in fair agreement with the experimental value[236] of -6.9 kcal mol^{-1}, particularly given the differences between the molecular system considered here and that studied experimentally.

As described in Section 4.2.2, not all of the hydrophobic pocket region of IQN17

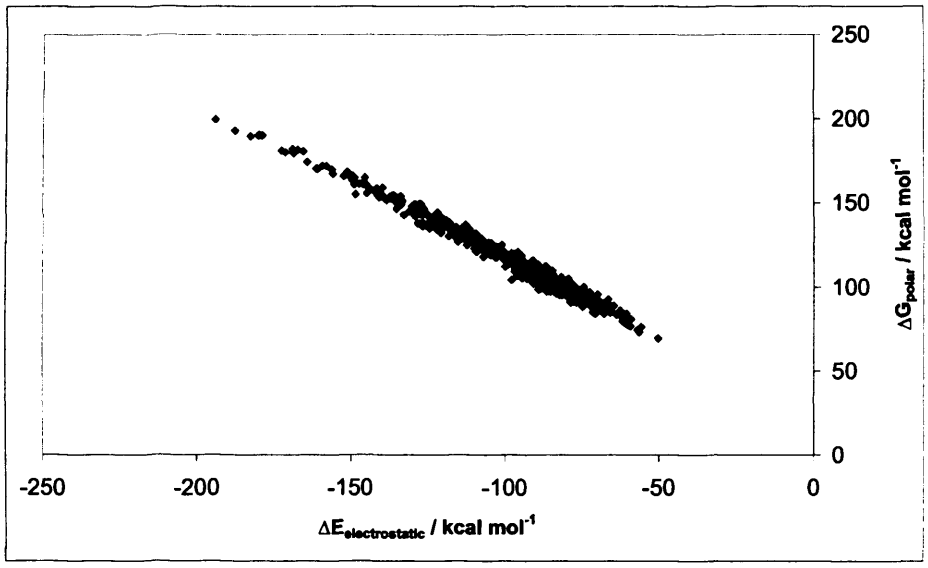

Figure 6.3: Plot of $\Delta E_{\text{electrostatic}}$ against ΔG_{polar} for the IQN17–D10-p1 complex. The correlation coefficient is -0.99.

is represented in the system used in this study, as some of the pocket residues come from the adjacent IQN17 helix. The effect of these missing residues on the binding energy must also be considered. The charged Lys-574 residue is the type of residue normally found on the surfaces of proteins but forms a salt bridge with Asp-632 in the C34 helix of the HIV-1 gp41 complex. No such salt bridge is formed in the IQN17–D10-p1 system. The presence of an intermolecular hydrogen bond between Gln-577 of the adjacent helix and Trp-12 of D10-p1 in the full trimeric system would lead to an increase in the magnitude of the binding energy compared to the system studied here. This will be discussed in more detail in the following section. The hydrophobic Leu-566 and Val-570 are typical

residues found in protein-protein interfaces and would increase the favourable intermolecular hydrophobic interactions, again increasing the magnitude of the binding energy. Therefore, it is reasonable to expect that the addition of the missing residues would lead to an increase in binding energy, in agreement with experimental observations.

6.1.2 Computational Alanine Scanning

Eight of the 16 residues of D10-p1 were mutated to alanine, as before, and the effect on the binding energy and its components are given in Table 6.4. These results are represented graphically in Figure 6.4. For obvious reasons, the five alanine residues were not mutated. The glycine residue was not mutated as the mutagenesis technique used here requires that, unlike glycine, the sidechain of the amino acid residue has both β- and γ-carbons. This represents one of the drawbacks of this method, as mutations to larger sidechains would be desirable for examining potential drugs. In this case, however, very little information would be gained from a mutation of the glycine residue to alanine. The two cysteine residues of D10-p1 were also not mutated due to their essential function in determining the circular nature of this class of D-peptide inhibitor.[21]

The residues Trp-10 and Leu-13 showed large, negative $\Delta\Delta G_{\text{subtotal}}$ upon mutation to alanine, indicating a highly unfavourable change in binding. The two glutamine residues, Glu-4 and Glu-9, showed small, negative $\Delta\Delta G_{\text{subtotal}}$. The residues Arg-6, His-7, Arg-8 and Trp-12 showed small, positive $\Delta\Delta G_{\text{subtotal}}$, indicating a slight preference for the alanine residue at these positions.

Alanine scanning experiments on the interhelical interactions in the HIV-1

	Mutation							
	Glu4Ala		**Arg6Ala**		**His7Ala**		**Arg8Ala**	
Contribution	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\Delta\Delta E_{\text{electroststic}}$	−87.0	10.3	57.2	3.7	−3.3	1.9	49.8	2.2
$\Delta\Delta E_{\text{vdW}}$	−0.3	0.1	0.0	0.0	−0.8	0.7	0.0	0.0
$\Delta\Delta E_{\text{gas}}$	−87.3	10.3	57.2	0.0	−4.1	2.4	49.8	2.2
$\Delta\Delta G_{\text{polar}}$	86.2	10.3	−56.4	3.6	4.5	2.5	−49.1	2.1
$\Delta\Delta G_{\text{non-polar}}$	0.0	0.0	0.0	0.0	−0.1	0.1	−0.01	0.0
$\Delta\Delta G_{\text{subtotal}}$	−1.1	0.2	0.7	0.1	0.3	0.3	0.64	0.0
	Glu9Ala		**Trp10Ala**		**Trp12Ala**		**Leu13Ala**	
Contribution	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\Delta\Delta E_{\text{electroststic}}$	−94.7	18.6	−0.2	2.1	−0.1	0.5	0.1	0.1
$\Delta\Delta E_{\text{vdW}}$	0.0	1.3	−8.0	1.0	−0.9	0.3	−4.7	0.7
$\Delta\Delta E_{\text{gas}}$	−94.7	18.0	−8.2	2.3	−1.0	0.6	−4.7	0.7
$\Delta\Delta G_{\text{polar}}$	93.8	16.8	4.5	2.0	1.4	0.5	1.2	0.2
$\Delta\Delta G_{\text{non-polar}}$	−0.2	0.1	−0.9	0.1	0.0	0.1	−0.4	0.1
$\Delta\Delta G_{\text{subtotal}}$	−1.0	2.0	−4.6	1.0	0.3	0.3	−3.9	0.7

Table 6.4: Computational alanine scanning mutagenesis energy changes and components / kcal mol^{-1} for the IQN17–D10-p1 system.

gp41 core by Wang *et al.*[237] have identified several key residues in the C34 helix that bind to the hydrophobic pocket region of N36. The residues Trp-628, Trp-631 and Ile-635 of the C-terminal helix all triggered a significant reduction in binding upon mutation to alanine. This consequently resulted in a reduction of the ability of gp41 to mediate cellular fusion. These key residues in the C34 helix of gp41 correspond in position to the residues Trp-10, Trp-12 and Leu-13, respectively, of D10-p1 in the complex with IQN17. However, due to the opposite chirality of the mirror image D-peptide, the orientations of the sidechains are different for Trp-12

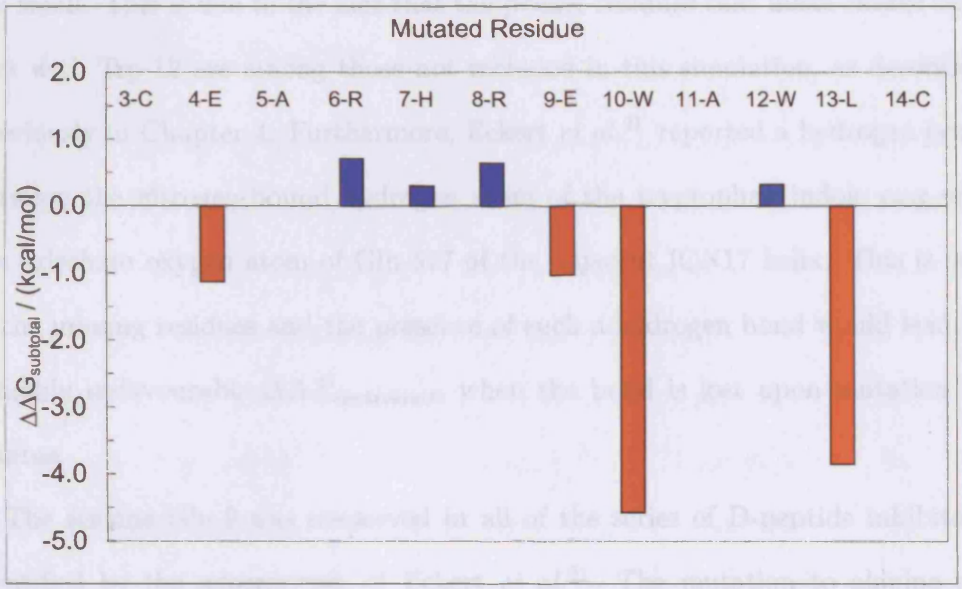

Figure 6.4: $\Delta\Delta G_{\text{subtotal}}$ / kcal mol^{-1} for the systematic mutation to alanine of each D-peptide residue in the IQN17–D10-p1 complex. Residues marked in blue indicate favourable mutations and residues marked in red indicate unfavourable mutations.

and Leu-13. The residue Trp-10 closely overlaps with the position of Trp-628 in the gp41 complex.

For Trp-10 and Leu-13, the calculated values of $\Delta\Delta G_{\text{subtotal}}$ are in agreement with the experimental observations.[237] These residues have bulky sidechains that participate in hydrophobic van der Waals interactions with the IQN17 pocket residues. These interactions are lost upon mutation to alanine and a large, negative $\Delta\Delta E_{\text{vdW}}$ results that is not compensated for by the gain in $\Delta\Delta G_{\text{polar}}$. The mutation of Trp-12 to alanine would also be expected to result in a similarly

large, negative $\Delta\Delta G_{\text{subtotal}}$. However, in this case, both $\Delta\Delta E_{\text{vdW}}$ and $\Delta\Delta G_{\text{polar}}$ are small. This is due to the fact that the pocket residues that make closest contact with Trp-12 are among those not included in this simulation, as described previously in Chapter 4. Furthermore, Eckert *et al.*[21] reported a hydrogen bond between the nitrogen-bound hydrogen atom of the tryptophan indole ring and the sidechain oxygen atom of Gln-577 of the adjacent IQN17 helix. This is one of the missing residues and the presence of such a hydrogen bond would lead to a highly unfavourable $\Delta\Delta E_{\text{electrostatic}}$ when the bond is lost upon mutation to alanine.

The residue Glu-9 was conserved in all of the series of D-peptide inhibitors identified by the experiments of Eckert *et al.*[21] The mutation to alanine results in a large, negative $\Delta\Delta E_{\text{electrostatic}}$ that is slightly larger than the opposing $\Delta\Delta G_{\text{polar}}$, leading to the observed small, negative $\Delta\Delta G_{\text{subtotal}}$. The same effect is also observed for the mutation of Glu-4 to alanine. The opposite effect is seen for mutation of the positively charged arginine residues, Arg-6 and Arg-8, as the favourable $\Delta\Delta E_{\text{electrostatic}}$ is slightly larger than the unfavourable $\Delta\Delta G_{\text{polar}}$. The positive $\Delta\Delta G_{\text{subtotal}}$ for the mutation of His-7 to alanine is negligible and is comparable to the standard deviation.

6.2 IQN17–C12 Interactions

6.2.1 Binding Energy

Figure 6.5 shows a plot of the RMSD in the positions of the α-carbons, relative to the X-ray crystal structure, for the data-gathering phase of the MD simulation of the IQN17–C12lm system. Figure 6.6 shows the same for the IQN17–C12unlm system. The C_α RMSD values are initially higher for the IQN17–C12unlm system than for the IQN17–C12lm system. This is likely to be related to the fact that the X-ray structure used for comparison is the IQN17–C12lm structure. Since the structure is less constrained, it is likely that the initial minimisation before the equilibration phase would lead to a greater deviation from the crystal structure than that seen for the cross-linked peptide. However, the standard deviations do not increase as the simulation takes its course and the simulations have similar C_α RMSD profiles to that seen for the IQN17–D10-p1 system. Once again, the net charge of $+2e$ is the likely cause of these fluctuations. The net charge of $+2e$ may contribute to this instability as systems with a net charge can exhibit artificial behaviour when PME is used, particularly under constant pressure conditions.[225] Furthermore, the attempts to limit the system size by using the smallest possible solvent box may also have contributed to the observed instability. As for the IQN17–D10-p1 system discussed previously, tt is likely that more stable simulations would have been achieved using counterions to make the system electronically neutral and using a larger solvent box to ensure that the entire system was surrounded by a significant amount of solvent at all times. The average values for C_α RMSD are 1.94 Å and 2.29 Å for the linked and un-

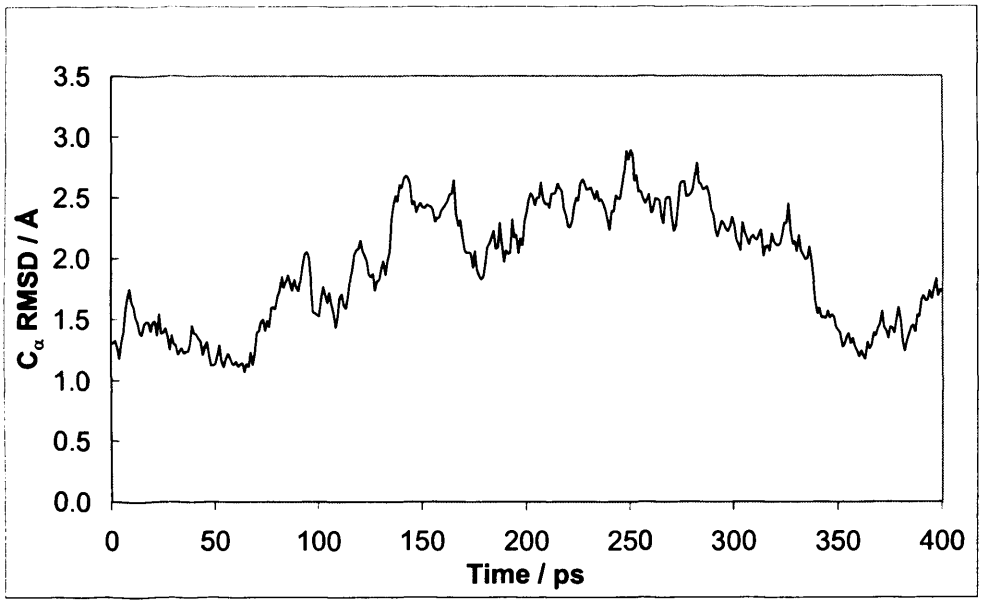

Figure 6.5: Root mean square deviation (RMSD) of the α-carbons in the IQN17–C12lm complex, relative to the X-ray crystal structure, during the data-gathering phase of the molecular dynamics trajectory.

linked systems, respectively. The respective standard deviations are 0.5 Å and 0.4 Å. These values are higher than for the relatively stable Mdm2–p53 systems discussed previously.

The absolute values for the energy and its components for each of the complex, protein and peptide systems are given in Table 6.5 for both IQN17–C12lm and IQN17–C12unlm. The entropy and its components are given in Table 6.6. As before, the structures have been minimised both with and without the application of the continuum solvation energy term. The binding free energy, $\Delta G_{\text{binding}}$, and

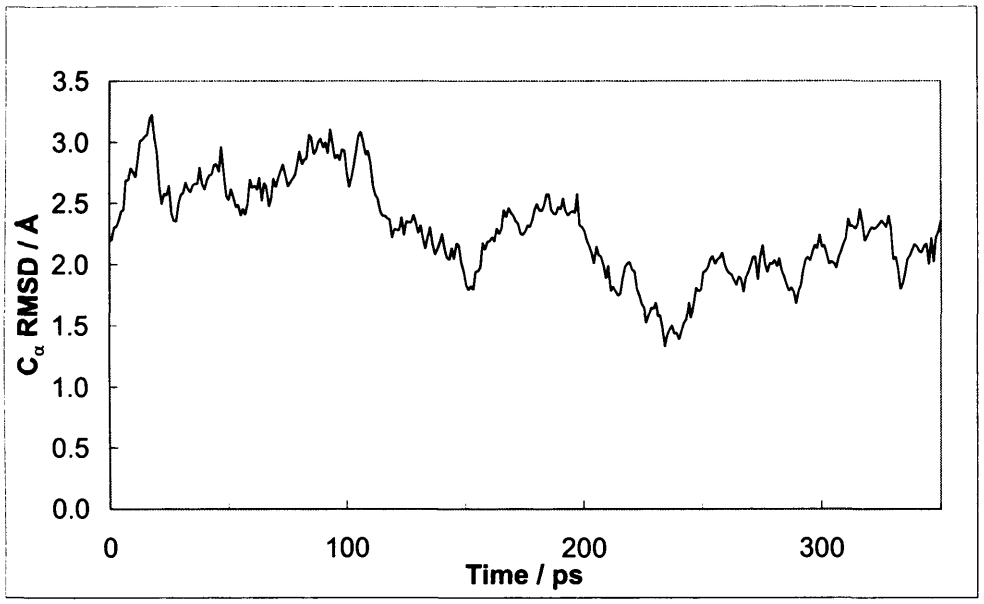

Figure 6.6: Root mean square deviation (RMSD) of the α-carbons in the IQN17–C12unlm complex, relative to the X-ray crystal structure, during the data-gathering phase of the molecular dynamics trajectory.

its components are given in Table 6.7. The given entropic contributions are given for structures minimised with the implicit solvent term applied.

The entropic contributions to the binding free energy, $-T\Delta S$, are estimated to be 26.8 and 29.1 kcal mol^{-1} for the linked and unlinked peptides, respectively, when the solvation term is not applied to the minimisation. For the solvated minimised structures, the values of $-T\Delta S$ are estimated to be 25.6 and 22.4 kcal mol^{-1}, respectively. As before, the structure minimised with the implicit solvation energy term was closer to the observed X-ray crystal structure than the

IQN17–C12lm						
	Complex		Protein		Peptide	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.
$E_{\text{electrostatic}}$	−1815.1	61.5	−1166.0	50.0	−466.4	20.2
E_{vdW}	−176.3	6.5	−108.2	6.0	−30.6	3.0
E_{internal}	1161.2	22.8	907.7	21.7	253.5	10.8
E_{gas}	−830.3	62.6	−366.6	53.7	−243.4	21.9
G_{polar}	−1665.1	52.1	−1395.5	41.1	−469.4	18.8
$G_{\text{non-polar}}$	44.8	0.7	39.1	0.6	12.2	0.4
G_{subtotal}	−2450.5	24.4	−1722.9	23.5	−700.6	10.4
IQN17–C12unlm						
	Complex		Protein		Peptide	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.
$E_{\text{electrostatic}}$	−1816.1	51.1	−1145.8	42.7	−474.6	17.4
E_{vdW}	−179.8	6.8	−107.9	6.1	−31.9	3.0
E_{internal}	1168.9	23.6	913.5	19.8	255.5	12.0
E_{gas}	−827.1	52.4	−340.2	42.2	−251.1	19.0
G_{polar}	−1655.8	45.6	−1412.6	38.1	−456.9	16.4
$G_{\text{non-polar}}$	45.3	0.8	39.2	0.6	13.0	0.5
G_{subtotal}	−2437.6	25.0	−1713.6	20.4	−694.9	12.2

Table 6.5: Absolute energy and components / kcal mol^{-1} for the IQN17–C12lm and IQN17–C12unlm complexes.

structure obtained by minimisation without the solvation term. Unsurprisingly, the IQN17–C12lm complex and peptide structures have C_α RMSD values that are smaller than for IQN17–C12unlm, given that the reference structure used is that of IQN17–C12lm. These C_α RMSD values have already been given in Chapter 4, in Table 4.8.

The experimentally determined values for $-T\Delta S$ are 6.2 and approximately 7

IQN17–C12lm						
	Complex		**Protein**		**Peptide**	
	No Solvent	Solvent	No Solvent	Solvent	No Solvent	Solvent
TS_{vib}	694.0	745.9	539.4	580.1	154.7	164.7
TS_{rot}	15.6	15.6	15.3	15.2	12.8	12.8
TS_{trans}	15.8	15.8	15.5	15.5	14.5	14.5
TS_{tot}	725.3	777.2	570.2	610.8	182.0	192.0
IQN17–C12unlm						
	Complex		**Protein**		**Peptide**	
	No Solvent	Solvent	No Solvent	Solvent	No Solvent	Solvent
TS_{vib}	693.3	756.0	539.4	580.1	156.2	171.6
TS_{rot}	15.5	15.6	15.3	15.2	12.8	12.8
TS_{trans}	15.8	15.8	15.5	15.5	14.5	14.5
TS_{tot}	724.6	787.4	570.2	610.8	183.5	198.9

Table 6.6: Entropy and components / kcal mol^{-1} for the IQN17–C12lm and IQN17–C12unlm complexes. Structures have been minimised with and without application of the continuum solvation term.

kcal mol^{-1} for the linked and unlinked peptides, respectively.[40] The calculations presented here overestimate the loss of entropy upon binding of the protein and peptide. Furthermore, the loss of entropy is expected to be less for the linked peptide, C12lm, as observed experimentally, yet this is not seen here for the solvent-minimised structures, although it is the case for the unsolvated minimised structures. Clearly, the method used here for estimating the entropy cannot adequately describe the conformational entropy of the C-peptide. This is perhaps not surprising given the assumptions inherent in the method. However, as stated previously, the loss of entropy upon binding is the most difficult contribution to

	IQN17–C12lm		IQN17–C12unlm	
Contribution	Mean	S.D.	Mean	S.D.
$\Delta E_{\text{electrostatic}}$	−182.8	20.2	−195.7	14.2
ΔE_{vdW}	−37.5	3.0	−40.1	3.4
ΔE_{gas}	−220.3	20.8	−235.8	15.4
ΔG_{polar}	199.8	20.1	213.6	14.0
$\Delta E_{\text{electrostatic}} + \Delta G_{\text{polar}}$	17.0	2.0	17.9	2.1
$\Delta G_{\text{non-polar}}$	−6.5	0.3	−6.9	0.4
$\Delta G_{\text{subtotal}}$	−27.1	2.7	−29.1	3.1
$-T\Delta S$ (solvent)	25.7	n/a	22.4	n/a
$\Delta G_{\text{binding}}$	−1.4	n/a	−6.7	n/a

Table 6.7: Binding energy and components / kcal mol^{-1} for the IQN17–C12lm and IQN17–C12unlm complexes.

quantify and it is not critical to do so for the computational alanine scanning studies. This relies on the assumption that the entropy change upon binding is essentially the same for two very similar mutants such as those studied here.

As seen for the systems studied previously, the contributions to the binding free energy arising from terms of electrostatic origin, $\Delta E_{\text{electrostatic}}$ and ΔG_{polar}, fluctuate considerably throughout the simulations for both linked and unlinked C-peptides. As before, the fluctuations tend to cancel one another out and the net effect, $\Delta E_{\text{electrostatic}} + \Delta G_{\text{polar}}$, has a much smaller standard deviation than either of the individual energy terms. This is illustrated graphically in Figures 6.7 and 6.9. This relationship is further illustrated in Figures 6.8 and 6.10, plots of $E_{\text{electrostatic}}$ against G_{polar}, which show the strong correlation between the two contributions. The calculated correlation coefficients were both −0.99, once again

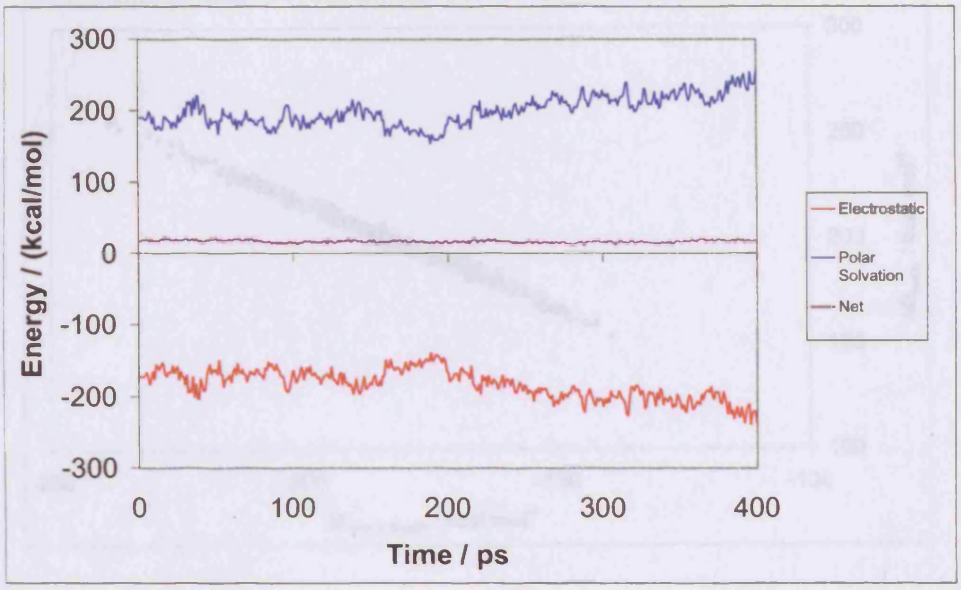

Figure 6.7: Contributions to the binding energy of the IQN17–C12lm complex due to electrostatic energy, polar solvation energy and the net effect of these two terms during the data-gathering phase of the molecular dynamics trajectory.

confirming the linear relationship.

The experimentally determined enthalpy change, ΔH, for the formation of both IQN17–C14lm and IQN17–C14unlm complexes is approximately -15 kcal mol^{-1} at 310 K.[40] The peptides used in this work would be expected to have similar values for $\Delta G_{\text{subtotal}}$ but the predicted values of -27 and -29 kcal mol^{-1} are somewhat larger than experiment. The predicted binding free energies, $\Delta G_{\text{binding}}$, of -1.4 and -6.7 kcal mol^{-1} for the linked and unlinked peptide, respectively, compare with experimental values of -8.4 and approximately -7 kcal mol^{-1}.[40]

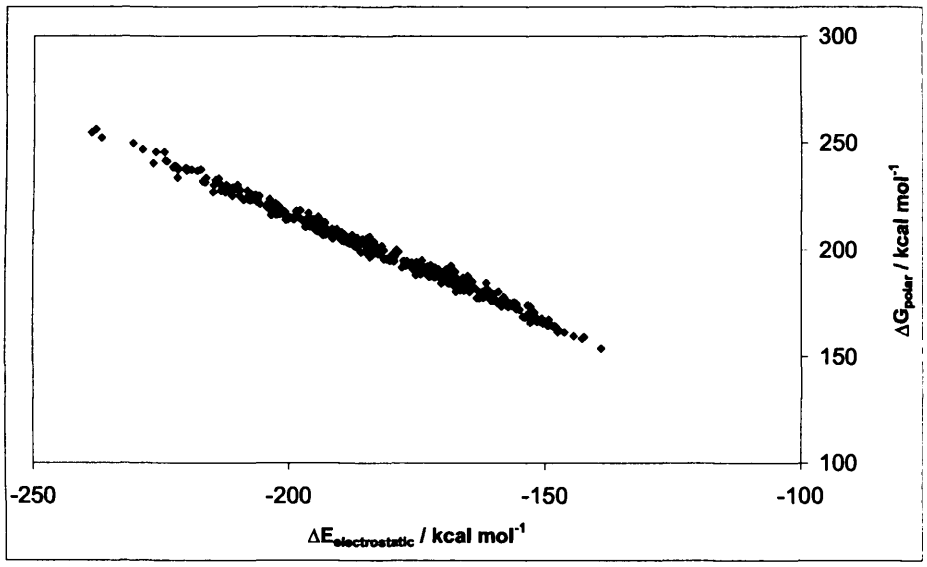

Figure 6.8: Plot of $\Delta E_{\text{electrostatic}}$ against ΔG_{polar} for the IQN17–C12lm complex. The correlation coefficient is -0.99.

The good agreement observed for the binding free energy of the IQN17–C12unlm complex arises from the overestimation of both the entropy and enthalpy of binding by similar amounts.

As for the IQN17–D10-p1 complex, the effect of the missing residues of the IQN17 hydrophobic pocket must be considered. Unlike for the D-peptide system, there are no hydrogen bonds between the missing residues and the C-peptides. This led to an underestimation of $\Delta G_{\text{subtotal}}$ for the D-peptide and partly accounts for the apparent overestimation of $\Delta G_{\text{subtotal}}$ seen here. As before, there would be a greater number of stabilising van der Waals interactions in the pocket region if all of the residues were included in the calculation. Furthermore, only 12 of

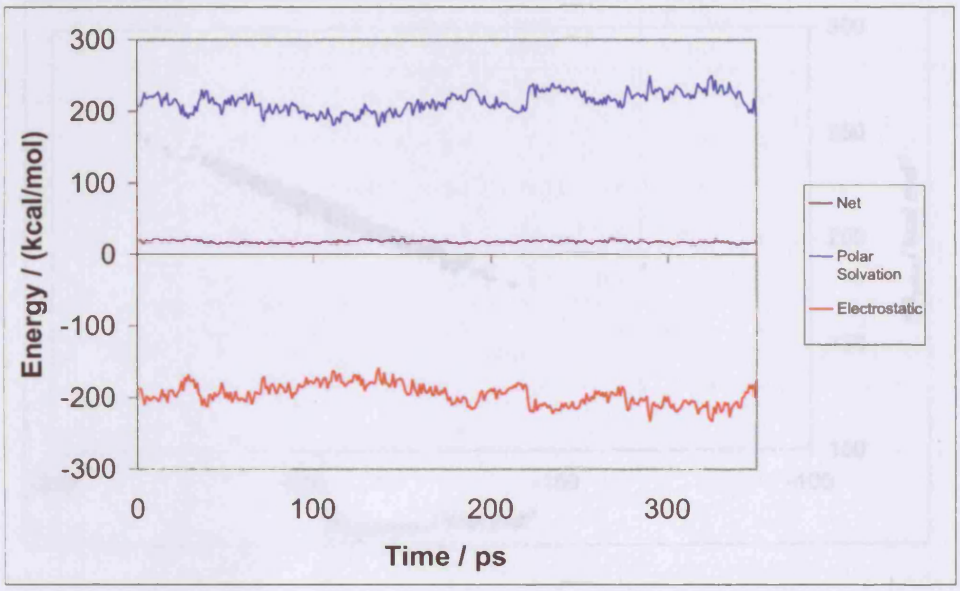

Figure 6.9: Contributions to the binding energy of the IQN17–C12unlm complex due to electrostatic energy, polar solvation energy and the net effect of these two terms during the data-gathering phase of the molecular dynamics trajectory.

the residues of the C14 sequence are included in the C12-peptides studied here.

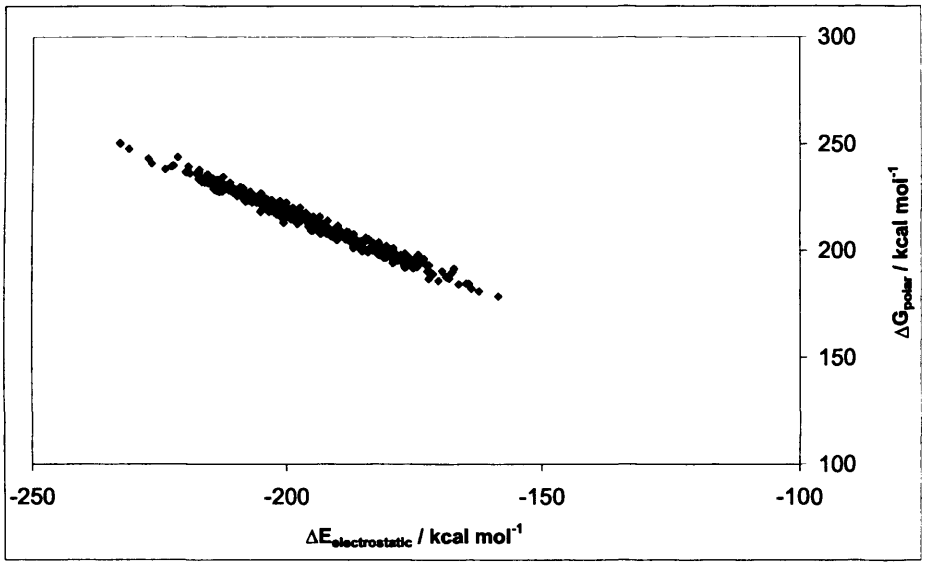

Figure 6.10: Plot of $E_{\text{electrostatic}}$ against G_{polar} for the IQN17–C12unlm complex. The correlation coefficient is -0.99.

6.2.2 Computational Alanine Scanning

Each residue of both C12lm and C12unlm has been mutated to alanine, with the exception of the two bridging glutamine residues, Gln-2 and Gln-9 (Gln-629 and Gln-646). The numbers in brackets correspond to the equivalent residue in gp41 taken from the HXB2 strain of HIV-1. The resulting effects on the binding energy and its components are given in Tables 6.8 and 6.9 for the IQN17–C12lm and IQN17–C12unlm complexes, respectively. The change in binding energy, $\Delta\Delta G_{\text{subtotal}}$, is shown for each residue in Figures 6.11 and 6.12, respectively, where a large, negative $\Delta\Delta G_{\text{subtotal}}$ indicates a highly unfavourable mutation, as

	Mutation									
	Trp1Ala		**Glu3Ala**		**Trp4Ala**		**Asp5Ala**		**Arg6Ala**	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
$\Delta\Delta E_{\text{electrostatic}}$	1.0	1.7	−60.8	3.2	−4.2	1.3	−65.7	2.7	60.4	2.5
$\Delta\Delta E_{\text{vdW}}$	−5.9	1.0	−0.1	0.0	−6.2	1.1	−0.1	0.0	0.0	0.0
$\Delta\Delta E_{\text{gas}}$	−4.9	2.1	−60.8	3.2	−10.4	1.2	−65.8	2.7	60.4	2.5
$\Delta\Delta G_{\text{polar}}$	2.3	1.7	60.4	3.2	5.5	0.9	65.4	2.7	−59.7	2.5
$\Delta\Delta G_{\text{non-polar}}$	−1.0	0.2	0.0	0.0	−0.5	0.1	0.0	0.0	0.00	0.0
$\Delta\Delta G_{\text{subtotal}}$	−3.7	1.0	−0.5	0.1	−5.5	0.9	−0.3	0.1	0.7	0.0
	Glu7Ala		**Ile8Ala**		**Asn10Ala**		**Tyr11Ala**		**Thr12Ala**	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
$\Delta\Delta E_{\text{electrostatic}}$	−94.0	11.2	−0.3	0.1	1.3	3.3	−2.9	2.5	0.6	0.4
$\Delta\Delta E_{\text{vdW}}$	−0.8	0.4	−3.2	0.6	−0.3	0.1	−6.5	1.0	−0.5	0.3
$\Delta\Delta E_{\text{gas}}$	−94.8	11.5	−3.5	0.6	−0.3	0.1	−9.5	2.4	0.1	0.5
$\Delta\Delta G_{\text{polar}}$	94.1	11.4	0.8	0.1	−1.1	3.4	4.7	2.1	0.4	0.7
$\Delta\Delta G_{\text{non-polar}}$	0.0	0.0	−0.2	0.1	0.01	0.0	−0.1	0.1	0.0	0.0
$\Delta\Delta G_{\text{subtotal}}$	−0.7	0.3	−2.8	0.6	−0.1	0.2	−5.4	1.0	0.4	0.4

Table 6.8: Computational alanine scanning mutagenesis energy changes and components / kcal mol^{-1} for the IQN17–C12lm system.

before.

The amino acid residues comprising the C12-peptides used in this work correspond to the stretch of residues from Trp-628 to Thr-639 of gp41 taken from the HXB2 strain of HIV-1.[40] Alanine scanning experiments on this stretch of residues identified four residues that contribute to the stability of the six-helical bundle, which is the fusion-active conformation of gp41.[237] Mutating the residues Trp-628, Trp-631 and Ile-635 to alanine led to a reduction in gp41-mediated fusion. The residue Tyr-638 also played a role in stabilising the six-helical bundle.[237]

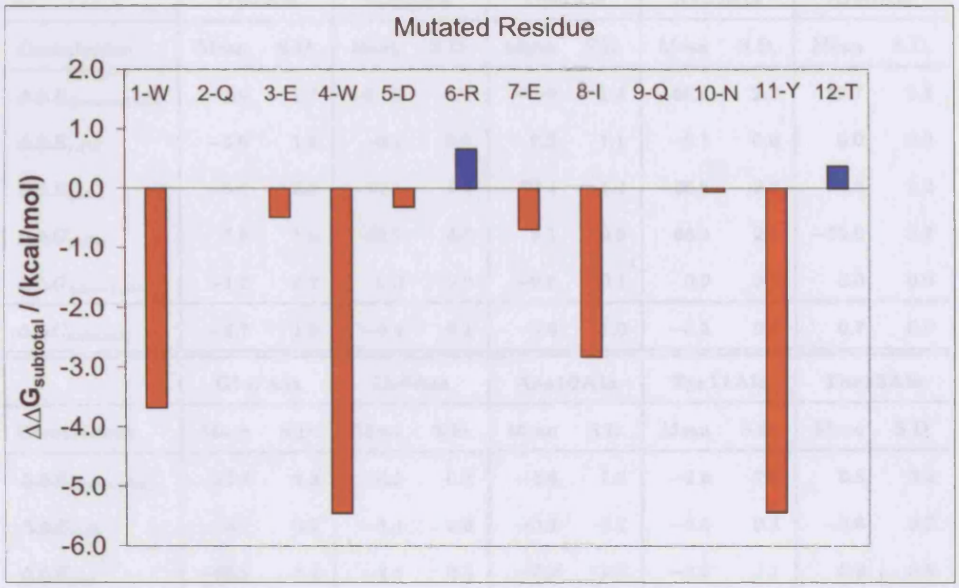

Figure 6.11: $\Delta\Delta G_{\text{subtotal}}$ / kcal mol^{-1} for the systematic mutation to alanine of each C-peptide residue in the IQN17–C12lm complex. Residues marked in blue indicate favourable mutations and residues marked in red indicate unfavourable mutations.

These residues correspond to Trp-1, Trp-4, Ile-8 and Tyr-11 of C12 as described in this work. Both numbering schemes are given for ease of comparison. X-ray crystallography studies have confirmed that these residues adopt positions similar to those seen for the corresponding residues of the C34 helix in the gp41 core.[40]

As would be expected, both IQN17–C12lm and IQN17–C12unlm complexes exhibit similar changes in binding energy upon systematic mutation of the peptide residues to alanine. All four residues identified by the alanine scanning experiments of Wang *et al.*[237] showed large, negative $\Delta\Delta G_{\text{subtotal}}$ here, indicating

	Mutation									
	Trp1Ala		**Glu3Ala**		**Trp4Ala**		**Asp5Ala**		**Arg6Ala**	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
$\Delta\Delta E_{\text{electrostatic}}$	−1.9	1.7	−63.0	4.1	−3.9	1.4	−66.6	2.3	59.7	2.3
$\Delta\Delta E_{\text{vdW}}$	−6.6	1.9	−0.1	0.0	−6.5	1.1	−0.1	0.0	0.0	0.0
$\Delta\Delta E_{\text{gas}}$	−8.5	2.8	−63.1	4.1	−10.4	1.2	−66.6	2.3	59.6	2.2
$\Delta\Delta G_{\text{polar}}$	5.0	1.8	62.7	4.0	5.1	0.9	66.3	2.3	−59.0	2.3
$\Delta\Delta G_{\text{non-polar}}$	−1.2	0.2	0.0	0.0	−0.6	0.1	0.0	0.0	0.0	0.0
$\Delta\Delta G_{\text{subtotal}}$	−4.7	1.9	−0.4	0.1	−5.9	1.0	−0.3	0.1	0.7	0.0
	Glu7Ala		**Ile8Ala**		**Asn10Ala**		**Tyr11Ala**		**Thr12Ala**	
Contribution	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
$\Delta\Delta E_{\text{electrostatic}}$	−92.3	8.3	−0.3	0.1	−4.6	1.6	−2.6	2.0	0.8	0.4
$\Delta\Delta E_{\text{vdW}}$	−0.7	0.2	−3.1	0.6	−0.2	0.1	−6.6	0.8	−0.6	0.2
$\Delta\Delta E_{\text{gas}}$	−93.1	8.4	−3.4	0.6	−4.9	1.7	−9.2	2.1	0.2	0.5
$\Delta\Delta G_{\text{polar}}$	92.4	8.4	0.8	0.1	5.2	1.7	4.8	1.7	0.1	0.8
$\Delta\Delta G_{\text{non-polar}}$	0.0	0.0	−0.2	0.0	0.0	0.0	−0.7	0.1	0.0	0.0
$\Delta\Delta G_{\text{subtotal}}$	−0.7	0.2	−2.7	0.6	0.3	0.2	−5.1	0.8	0.3	0.4

Table 6.9: Computational alanine scanning mutagenesis energy changes and components / kcal mol^{-1} for the IQN17–C12unlm system.

highly unfavourable substitutions. The remaining residues showed small changes in binding energy upon mutation, with Arg-6 (Arg-633) showing the greatest preference for alanine at any of the peptide positions.

In the cases of the mutation of Ala-1 (Ala-628), Ile-8 (Ile-635) and Tyr-11 (Tyr-638) to alanine the main contribution to the observed negative $\Delta\Delta G_{\text{subtotal}}$ arises from the loss of stabilising van der Waals interactions between the bulky sidechains and the binding site following the mutation. This results in a large, negative $\Delta\Delta E_{\text{vdW}}$ that is not balanced by the net effect of the terms of electro-

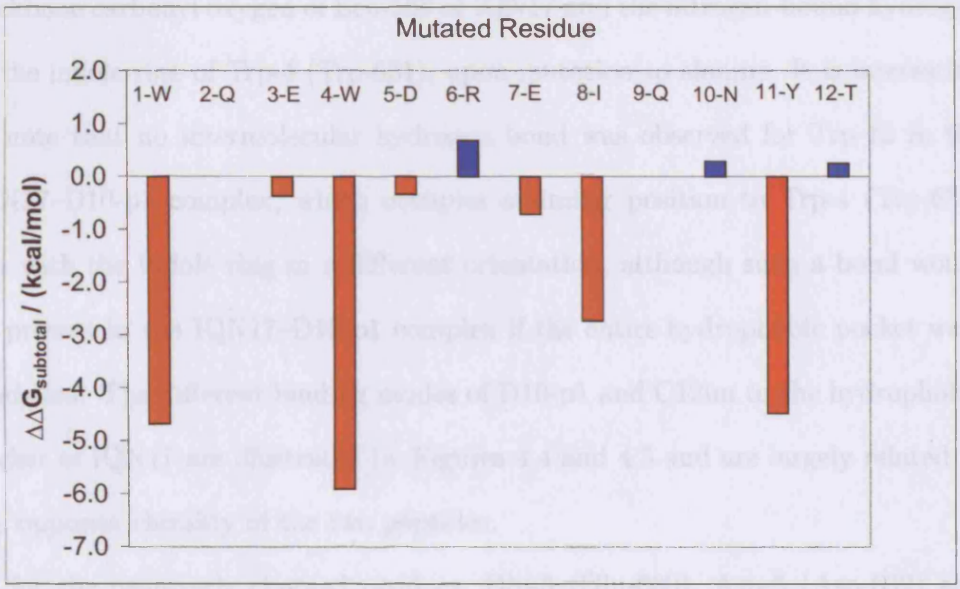

Figure 6.12: $\Delta\Delta G_{\text{subtotal}}$ / kcal mol^{-1} for the systematic mutation to alanine of each C-peptide residue in the IQN17–C12unlm complex. Residues marked in blue indicate favourable mutations and residues marked in red indicate unfavourable mutations.

static origin, $\Delta\Delta E_{\text{electrostatic}} + \Delta\Delta G_{\text{polar}}$. In the case of the Tyr11Ala (Tyr638Ala) mutation there is an electrostatic contribution of approximately -3 kcal mol^{-1} opposing the mutation. This is not strong enough to suggest a lost hydrogen bond upon mutation and no such bond is observed in the X-ray crystal structure of IQN17–C12lm.

For the Trp-4 (Trp-631) mutation to alanine there is a very large, negative $\Delta\Delta G_{\text{subtotal}}$. In part, this can be accounted for in a similar fashion to the previous mutations. However, there is also a contribution of -4.2 kcal mol^{-1} from

$\Delta\Delta E_{\text{electrostatic}}$ due to the loss of an intermolecular hydrogen bond, between the backbone carbonyl oxygen of Leu-568 of IQN17 and the nitrogen-bound hydrogen of the indole ring of Trp-4 (Trp-631), upon mutation to alanine. It is interesting to note that no intermolecular hydrogen bond was observed for Trp-12 in the IQN17–D10-p1 complex, which occupies a similar position to Trp-4 (Trp-631) but with the indole ring in a different orientation, although such a bond would be present in the IQN17–D10-p1 complex if the entire hydrophobic pocket were modelled. The different binding modes of D10-p1 and C12lm to the hydrophobic pocket of IQN17 are illustrated by Figures 4.4 and 4.5 and are largely related to the opposite chirality of the two peptides.

For the negatively charged residues, Glu-3 (Glu-630), Asp-5 (Asp-632) and Glu-7 (Glu-634), the large, negative $\Delta\Delta E_{\text{electrostatic}}$ opposing the mutation is balanced by the large, positive $\Delta\Delta G_{\text{polar}}$ and the net effect is small. Conversely, the positively charged residue Arg-6 (Arg-633) has a positive $\Delta\Delta E_{\text{electrostatic}}$ and a negative $\Delta\Delta G_{\text{polar}}$ upon mutation to alanine but the net effect is still small. These sidechains are also relatively small or are situated far from the hydrophobic binding site and the resulting loss of van der Waals interactions is small for each mutation.

Part IV

Conclusions and Further Work

Chapter 7

Conclusions

The computational alanine scanning technique first implemented by Massova and Kollman[3,174] has been confirmed as a valuable tool for the investigation of protein–peptide binding interfaces. This technique has the advantage that important information on the key factors in the binding of two polypeptide molecules can be obtained without having to calculate the entropic contributions to the binding free energy. Treating the entropy using classical ideal gas thermodynamics has been shown to give entropic contributions of the correct order of magnitude.

The technique has been applied to the previously studied interaction between the oncoprotein Mdm2 and the transcription regulator and tumour suppressor peptide p53, both for *Xenopus Laevis* and human systems. Qualitative agreement has been observed between this work and previous theoretical and experimental investigations. The key residues in the binding interface have been identified in agreement with amino acid replacement studies and quantitative information

regarding the contributions to the binding energy of each of these residues has been obtained. In most cases, van der Waals interactions in the binding interface have been found to be of greater importance than any net electrostatic effects, although an intermolecular hydrogen bond has been identified between a tryptophan residue and the largely hydrophobic binding site.

The technique has also been extended to the interactions of several potential inhibitors of the gp41-mediated viral fusion process in HIV-1 infection. Computational alanine scanning mutagenesis has been used to identify the key residues in the binding of a D-peptide and two similar C-peptides to IQN17, a model of a transiently exposed potential drug target in the fusion process. Good agreement with experimental alanine scanning studies on the HIV gp41 core was seen for most of the key residues in the D-peptide and the only discrepancy has been attributed to the incompleteness of the protein structure used, necessitated by the computational expense of modelling the complete trimeric system. Good agreement with the alanine scanning experiments has been observed for all of the key residues of the C-peptide inhibitors. Once again, van der Waals interactions have been found to play a key role in the binding interface and an intermolecular hydrogen bond that contributes to the stability of the complex has also been identified.

This technique allows the contributions of the key residues to the binding free energy to be decomposed according to their physical origin, an advantage over the analogous experimental methods. The agreement with experiments for all of the systems studied is encouraging, suggesting that the systematic mutation of residues, as implemented here, is a useful tool for predicting the relative activity

of peptides as drug candidates. The main computational cost of such studies is invested in determining the molecular dynamics trajectories and with the advances in algorithms, increasing parallelisation and computational power, this method could provide a fast and effective way of screening proteins and peptides for such inhibitory activity.

Chapter 8

Future Work

The computational alanine scanning technique utilised in this work is completely general and could readily be applied to any non-covalent protein–peptide interaction for which a suitable molecular mechanical force field was available, with the only limits being those of computational resources. For example, enveloped viruses with similar cellular fusion mechanisms to HIV-1, such as influenza, could be studied using the method in order to identify the key binding residues. It would be relatively simple to alter the MDANAL code developed during this work to allow mutation of residues in the binding site of the target protein, giving further insights into the factors affecting protein–peptide binding.

As computational resources and the use of efficient parallelisation techniques increase, larger systems will be accessible using similar approaches. For example, this would allow simulation of the entire trimeric core of HIV-1 gp41, allowing the entire hydrophobic pocket region to be modelled. It would be of interest to study a range of D-peptide inhibitors in order to study the link between binding

free energy and strength of inhibitory activity. It would also be of interest to study larger C-peptides such as C34 in complex with the N36 trimeric core in order to investigate the interactions along the hydrophobic groove on the surface of the trimer, as well as the more highly-conserved pocket region.

The most approximate part of this work is the treatment of entropy using classical ideal gas thermodynamics in the evaluation of the binding free energy. It would be desirable to improve the treatment, possibly by explicitly considering the conformational entropy of the backbone and sidechain groups.

Appendix A

Amino Acids

A.1 Identification Codes

Name			Sidechain
Alanine	Ala	A	$-CH_3$
Valine	Val	V	$-CH(CH_3)_2$
Leucine	Leu	L	$-CH_2CH(CH_3)_2$
Isoleucine	Ile	I	$-CH(CH_3)CH_2CH_3$

Table A.1: The hydrophobic aliphatic amino acid sidechains and identification codes.

Name			Sidechain	Notes
Phenylalanine	Phe	F	$-CH_2C_6H_5$	
Tryptophan	Trp	W	$-CH_2$IND	IND = indole ring
Tyrosine	Tyr	Y	$-CH_2C_6H_4OH$	phenyl ring is 1,4 substituted

Table A.2: The hydrophobic aromatic amino acid sidechains and identification codes.

Name			Sidechain	Notes
Aspartic Acid	Asp	D	$-CH_2COO^-$	
Glutamic Acid	Glu	E	$-CH_2CH_2COO^-$	
Lysine	Lys	K	$-CH_2CH_2CH_2CH_2NH_3^+$	
Arginine	Arg	R	$-CH_2CH_2CH_2NHC(NH_2) = NH_2^+$	
Histidine	His	H	$-CH_2\mathbf{C} = CHN = \mathbf{CH}NH$	ring; uncharged at physiological pH

Table A.3: The charged amino acid sidechains and identification codes.

Name			Sidechain	Notes
Serine	Ser	S	$-CH_2OH$	
Threonine	Thr	T	$-CH(OH)CH_3$	
Cysteine	Cys	C	$-CH_2SH$	largely non-polar; can form disulphide bridges
Methionine	Met	M	$-CH_2CH_2SCH_3$	largely non-polar
Asparagine	Asn	N	$-CH_2CONH_2$	
Glutamine	Gln	Q	$-CH_2CH_2CONH_2$	

Table A.4: The neutral polar amino acid sidechains and identification codes.

Name			Sidechain	Notes
Glycine	Gly	G	-H	achiral
Proline	Pro	P	$-CH_2CH_2\mathbf{C}H_2$	**C** bound to backbone N to form ring

Table A.5: The conformationally important amino acid sidechains and identification codes.

Appendix B

Entropy Codes

B.1 ROTRANSENT

```
c
c
c     ######################################################################
c     ##                                                                  ##
c     ## program rotransent  --  rotational and translational entropy    ##
c     ##                                                                  ##
c     ######################################################################
c
c
c     "rotransent" calculates the total mass and principal moments of
c     inertia of the molecule and subsequently calculates the rotational
c     and translational entropy of the system
c
c     NB: The code for calculating the moments of inertia were taken
c     from the TINKER package
c
c     Literature Reference :
c
c     D. A. McQuarrie, "Statistical Mechanics", Harper and Row,
c     New York, 1976
c
c
      program rotransent
```

```
      implicit none
      include 'sizes.i'
      include 'atoms.i'
      include 'atmtyp.i'
      include 'iounit.i'
      include 'math.i'
      include 'units.i'
      integer i,j,k
      real*8 tstrans,tsrot,ztrans,zrot
      real*8 kb,pl,na
      real*8 temp,pres
      real*8 ia,ib,ic
      real*8 weigh,total,dot
      real*8 xcm,ycm,zcm
      real*8 xx,xy,xz,yy,yz,zz
      real*8 xterm,yterm,zterm
      real*8 moment(3),vec(3,3)
      real*8 work1(3),work2(3)
      real*8 tensor(3,3),a(3,3)
c
c
c     set up the structure and mechanics calculation
c
      call initial
      call getxyz
      call mechanic
c
c     calculate the position of the centre of mass
c
      total = 0.0d0
      xcm = 0.0d0
      ycm = 0.0d0
      zcm = 0.0d0
      do i = 1, n
         weigh = mass(i)
         total = total + weigh
         xcm = xcm + x(i)*weigh
         ycm = ycm + y(i)*weigh
         zcm = zcm + z(i)*weigh
      end do
      xcm = xcm / total
      ycm = ycm / total
      zcm = zcm / total
c
c     print out the total mass
```

```
c
      write (iout,5) total
 5    format (/,' Total Mass (amu) : ',f16.2)
c
c     compute and then diagonalize the inertial tensor
c
      xx = 0.0d0
      xy = 0.0d0
      xz = 0.0d0
      yy = 0.0d0
      yz = 0.0d0
      zz = 0.0d0
      do i = 1, n
         weigh = mass(i)
         xterm = x(i) - xcm
         yterm = y(i) - ycm
         zterm = z(i) - zcm
         xx = xx + xterm*xterm*weigh
         xy = xy + xterm*yterm*weigh
         xz = xz + xterm*zterm*weigh
         yy = yy + yterm*yterm*weigh
         yz = yz + yterm*zterm*weigh
         zz = zz + zterm*zterm*weigh
      end do
      tensor(1,1) = yy + zz
      tensor(2,1) = -xy
      tensor(3,1) = -xz
      tensor(1,2) = -xy
      tensor(2,2) = xx + zz
      tensor(3,2) = -yz
      tensor(1,3) = -xz
      tensor(2,3) = -yz
      tensor(3,3) = xx + yy
      call jacobi (3,3,tensor,moment,vec,work1,work2)
c
c     select the direction for each principal moment axis
c
      do i = 1, 2
         do j = 1, n
            xterm = vec(1,i) * (x(j)-xcm)
            yterm = vec(2,i) * (y(j)-ycm)
            zterm = vec(3,i) * (z(j)-zcm)
            dot = xterm + yterm + zterm
            if (dot .lt. 0.0d0) then
               do k = 1, 3
```

```
                  vec(k,i) = -vec(k,i)
               end do
            end if
            if (dot .ne. 0.0d0)  goto 10
         end do
   10    continue
      end do
c
c     moment axes must give a right-handed coordinate system
c
      xterm = vec(1,1) * (vec(2,2)*vec(3,3)-vec(2,3)*vec(3,2))
      yterm = vec(2,1) * (vec(1,3)*vec(3,2)-vec(1,2)*vec(3,3))
      zterm = vec(3,1) * (vec(1,2)*vec(2,3)-vec(1,3)*vec(2,2))
      dot = xterm + yterm + zterm
      if (dot .lt. 0.0d0) then
         do j = 1, 3
            vec(j,3) = -vec(j,3)
         end do
      end if
c
c     print the moments of inertia and the principal axes
c
      write (iout,20)
 20   format (/,' Moments of Inertia and Principal Axes :',
     &     //,9x,'Moment',15x,'X-, Y- and Z-Components of Axis')
      write (iout,30)  (moment(i),vec(1,i),vec(2,i),vec(3,i),i=1,3)
 30   format (3(/,f16.4,9x,3f12.6))
c
c     set some SI constants
c
      kb = 1.38066d-23
      pl = 6.62608d-34
      na = 6.02214d+23
      temp = 300.0d0
      pres = 1.0d+5
c
c     zero the partition function and entropy values
c
      zrot    = 0.0d0
      ztrans  = 0.0d0
      tsrot   = 0.0d0
      tstrans = 0.0d0
c
c     calculate the translational partition function
c
```

```
      total = total * 1.66054d-27
      ztrans = sqrt((2*pi*total*kb*temp/pl**2)**3) * (kb*temp/pres)
      write (iout,40) ztrans
 40   format (/,' Translational Partition Function = ',e16.6)
c
c     calculate the rotational partition function, assuming that sigma,
c     the symmetry number, is always unity for biomolecules
c
      ia = moment(1) * 1.66054d-47
      ib = moment(2) * 1.66054d-47
      ic = moment(3) * 1.66054d-47
      zrot = sqrt(pi*ia*ib*ic) * sqrt((8*(pi**2)*kb*temp/(pl**2))**3)
      zrot = 0.5*zrot
      write (iout,50) zrot
 50   format (/,' Rotational Partition Function    = ',e16.6)
c
c     calculate the translational entropy (TS) from the translational
c     partition function
c
      tstrans = (2.5 + log(ztrans))*na*kb*temp
      tstrans = tstrans / 4184
c
c     calculate the rotational entropy (TS) from the rotational
c     partition function
c
      tsrot = (1.5 + log(zrot))*na*kb*temp
      tsrot = tsrot / 4184
c
c     print out the values for translational and rotational entropy (TS)
c
      write (iout,60) tstrans
 60   format (/,' Translational Entropy (kcal/mol) = ',f16.6)
      write (iout,70) tsrot
 70   format (/,' Rotational Entropy (kcal/mol)    = ',f16.6)
c
c     end the program
c
      end
```

B.2 VIBENT

```
c
c
c     ################################################################
c     ##                                                            ##
c     ## program vibent  --  normal modes and vibrational entropy   ##
c     ##                                                            ##
c     ################################################################
c
c
c     "vibent" performs a vibrational normal mode analysis as in the
c     program "vibrate", followed by a calculation of the vibrational
c     entropy of the system
c
c     NB: the code for the normal mode analysis was taken from the
c     TINKER package
c
c
      program vibent
      implicit none
      include 'sizes.i'
      include 'atmtyp.i'
      include 'atoms.i'
      include 'files.i'
      include 'hescut.i'
      include 'iounit.i'
      include 'math.i'
      include 'units.i'
      integer i,j,k,m,ixyz,ihess,lext,freeunit
      integer nfreq,ndummy,nvib,ivib,nview
      integer hinit(3,maxatm),hstop(3,maxatm),hindex(maxhess)
      real*8 pl,kb,na,temp,light
      real*8 tsvib1,tsvib2,tsvib3,tsvib,tsvibtot
      real*8 h(maxhess),hdiag(3,maxatm)
      real*8 xref(maxatm),yref(maxatm),zref(maxatm)
      real*8 mass2(maxatm),factor,vnorm,ratio
      real*8 matrix((maxvib+1)*maxvib/2)
      real*8 eigen(maxvib),vects(maxvib,maxvib)
      real*8 a(maxvib+1),b(maxvib+1),p(maxvib+1),w(maxvib+1)
      real*8 ta(maxvib+1),tb(maxvib+1),ty(maxvib+1)
      character*1 eigunits(maxvib)
      character*7 ext
      character*60 coordfile,string
      logical exist,query
```

```
c
c
c     set up the structure and mechanics calculation
c
      call initial
      call getxyz
      call mechanic
c
c     initialize various things needed for vibrations
c
      nfreq = 3 * n
      ndummy = 0
      if (nfreq.gt.maxvib .or. nfreq**2.gt.maxhess) then
         write (iout,10)
   10    format (/,' VIBRATE  --  Too many Atoms in the Molecule')
         call fatal
      end if
      do i = 1, n
         if (atomic(i) .eq. 0) then
            ndummy = ndummy + 1
            mass(i) = 0.001d0
         end if
         mass2(i) = sqrt(mass(i))
      end do
      nvib = nfreq - 3*ndummy
c
c     calculate the Hessian matrix of second derivatives
c
      hesscut = 0.0d0
      call hessian (h,hinit,hstop,hindex,hdiag)
c
c     store upper triangle of the Hessian in "matrix"
c
      ihess = 0
      do i = 1, n
         do j = 1, 3
            ihess = ihess + 1
            matrix(ihess) = hdiag(j,i)
            do k = hinit(j,i), hstop(j,i)
               ihess = ihess + 1
               matrix(ihess) = h(k)
            end do
         end do
      end do
c
```

```
c     perform diagonalization to get Hessian eigenvalues
c
      call diagq (nfreq,maxvib,nfreq,matrix,eigen,vects,
     &                    a,b,p,w,ta,tb,ty)
      write (iout,20)
   20 format (/,' Eigenvalues of the Hessian Matrix :',/)
      write (iout,30)  (i,eigen(i),i=1,nvib)
   30 format (5(i5,f9.3,1x))
      write (iout,40)
   40 format ()
c
c     store upper triangle of the mass-weighted Hessian matrix
c
      ihess = 0
      do i = 1, n
         do j = 1, 3
            ihess = ihess + 1
            matrix(ihess) = hdiag(j,i) / mass(i)
            do k = hinit(j,i), hstop(j,i)
               m = (hindex(k)+2) / 3
               ihess = ihess + 1
               matrix(ihess) = h(k) / (mass2(i)*mass2(m))
            end do
         end do
      end do
c
c     diagonalize to get vibrational frequencies and normal modes
c
      call diagq (nfreq,maxvib,nfreq,matrix,eigen,vects,
     &                    a,b,p,w,ta,tb,ty)
      factor = sqrt(convert) / (2.0d0*pi*lightspd)
      do i = 1, nvib
         if (eigen(i) .lt. 0) then
            eigunits(i) = 'I'
         else
            eigunits(i) = ' '
         end if
         eigen(i) = factor * sqrt(abs(eigen(i)))
      end do
      write (iout,50)
   50 format (/,' Vibrational Frequencies (cm-1) :',/)
      write (iout,60)  (i,eigen(i),eigunits(i),i=1,nvib)
   60 format (5(i5,f9.3,a1))
c
c     TEST FOR PROPANE FREQUENCIES
```

```
c
c       eigen(1)  = 0.0d0
c       eigen(2)  = 0.0d0
c       eigen(3)  = 0.0d0
c       eigen(4)  = 0.0d0
c       eigen(5)  = 0.0d0
c       eigen(6)  = 0.0d0
c       eigen(7)  = 375.0d0
c       eigen(8)  = 748.0d0
c       eigen(9)  = 868.0d0
c       eigen(10) = 922.0d0
c       eigen(11) = 940.0d0
c       eigen(12) = 1053.0d0
c       eigen(13) = 1155.0d0
c       eigen(14) = 1179.0d0
c       eigen(15) = 1278.0d0
c       eigen(16) = 1338.0d0
c       eigen(17) = 1370.0d0
c       eigen(18) = 1375.0d0
c       eigen(19) = 1450.0d0
c       eigen(20) = 1450.0d0
c       eigen(21) = 1460.0d0
c       eigen(22) = 1468.0d0
c       eigen(23) = 1470.0d0
c       eigen(24) = 2914.0d0
c       eigen(25) = 2942.0d0
c       eigen(26) = 2960.0d0
c       eigen(27) = 2966.0d0
c       eigen(28) = 2968.0d0
c       eigen(29) = 2968.0d0
c       eigen(30) = 2970.0d0
c       eigen(31) = 2980.0d0
c       eigen(32) = 202.0d0
c       eigen(33) = 283.0d0
c
c      set some SI constants
c
c       kb = boltzmann
c       pl = planck
c       na = avogadro
c       temp = 300.0

        kb     = 1.38066d-23
        pl     = 6.62608d-34
        na     = 6.02214d+23
```

```
      light = 2.99792d+8
      temp  = 3.00d+2
c
c     calculate the vibrational entropy (TS) in kcal/mol
c
c     note that the factor of 100 in tsvib1 converts eigen(j) from
c     cm^-1 to m^-1
c
      tsvibtot = 0.0d0
      tsvib = 0.0d0
      do j = 7, nvib
         tsvib1 = 100*pl*light*eigen(j)/(kb*temp)
         tsvib2 = exp(tsvib1) - 1
         tsvib3 = 1 - exp(-tsvib1)
         tsvib  = ((tsvib1/tsvib2) - (log(tsvib3)))*na*kb*temp
         tsvib  = tsvib / 4184
         tsvibtot = tsvibtot + tsvib
         write (iout,70) j,eigen(j),tsvib
 70      format (/,i4,f16.6,f16.6)
      end do
      write (iout,*)
      write (iout,80) tsvibtot
 80   format (/,' Total Vibrational Entropy (kcal/mol) : ',f16.6)
c
c     perform any final tasks before program exit
c
      call final
      end
```

Appendix C

List of Abbreviations

ABP	arabinose binding protein
AIDS	acquired immunodeficiency syndrome
AspRS	aspartyl tRNA synthetase
ATP	adenosine triphosphate
BPTI	bovine pancreatic trypsin inhibitor
C_α, C_β etc.	α-carbon, β-carbon etc.
CK2	casein kinase 2
CMC/MD	chemical Monte Carlo/molecular dynamics
cyto	cytoplasmic
D	dextrorotatory
DNA	deoxyribonucleic acid

eq	equilibrium
ER	estrogen receptor
(US) FDA	(US) Food and Drug Administration
FEP	free energy perturbation
FFT	fast fourier transform
fp	fusion peptide
GB(SA)	generalised Born (surface area)
GTP	guanine triphosphate
HCV	hepatitis C virus
HDAC	histone deacetylase
HEI	hydroxyethylene inhibitor
HGH	human growth hormone
HIV-1	human immunodefiency virus type 1
HRSV	human respiratory syncytial virus
IC_{50}	inhibitory concentration to reduce activity by 50%
IDV	Indinavir
IR	infrared
LIE	linear interaction energy

M	molar, mol dm^{-3}
MBP	mannose binding protein
MC	Monte Carlo
MD	molecular dynamics
MM	molecular mechanics, molecular mechanical
MMLV	Moloney murine leukaemia virus
NMR	nuclear magnetic resonance
OWFEG	one-window free energy grid
PB(SA)	Poisson-Boltzmann (surface area)
PBC	periodic boundary conditions
PDB	protein data bank
PES	potential energy surface
PME	particle mesh Ewald
PI	path integral, predictive index
(HIV-)PR	(HIV) protease
PROFEC	pictorial representation of free energy components
QM	quantum mechanics, quantum mechanical

RCSB	Research Collaboratory for Structural Bioinformatics
RMS(D)	root mean square (deviation)
RNA	ribonucleic acid
(HIV-)RT	(HIV) reverse transcriptase
(SA)SA	(solvent accessible) surface area
SQV	Saquinavir
T	thymine
TI	thermodynamic integration
tm	transmembrane
TS	thymidylate synthetase
TyrRS	tyrosyl tRNA synthetase
U	deoxyuracil
UV	ultraviolet
vdW	van der Waals
WHAM	weighted histogram analysis method
WNDR	weighted non-polar desolvation ratio
X. Laevis	*Xenopus Laevis*, African clawed frog

Appendix D

List of Symbols

A	Helmholtz free energy
$\mathbf{a}$	acceleration
α	electrostatic coefficient in LIE
α	effective Born radius
α	constant governing width of Gaussian in PME
β	hydrophobic coefficient in LIE
Δ	change in
δt	time step
E	energy
$\mathrm{erfc}(x)$	complementary error function of x
ϵ	permittivity, dielectric constant

$\mathbf{F}$	force
f_{GB}	generalised Born function
G	Gibbs free energy
$\mathbf{G}$	constraint force
$\mathbf{g}$	gradient vector
H	enthalpy
$\mathscr{H}$	Hamiltonian
$\mathbf{H}$	Hessian matrix
h	Planck constant
$\hbar$	$h/2\pi$
I	moment of inertia
K	kinetic energy
K	equilibrium constant
K_r, K_θ	stretching, bending force constants
κ	Debye-Hückel parameter
k	Boltzmann constant
L	length of a cubic simulation box
λ	coupling parameter in FEP/TI
λ	multiplier used in geometric constraints
λ	multiplier used for velocity scaling in constant temperature MD

M	number of simulation steps
m	mass
$\mathbf{m}$	reciprocal space vector in PME
μ	multiplier used for length scaling in constant pressure MD
N	number of particles
$\mathbf{n}$	coordinate space vector in PME
ν	normal mode frequency
∇	nabla, differential operator
P	pressure
$\mathbf{p}$	momentum
Φ	electrostatic potential
ϕ	torsional angle
Ψ	total wavefunction
ψ	electronic wavefunction
ψ	torsional angle
$\mathbf{Q}$	mass-wieghted displacement coordinates
q	charge

R	molar gas constant
$\mathbf{R}$	position vector of atomic nuclei
$\mathbf{r}$	position vector of atoms or electrons, dependent on context
r	bond length
ρ	charge density
ρ	probability density
S	entropy
$S(r)$	truncation function
$S(\mathbf{m})$	structure factor in PME
σ	atomic solvation parameter
σ	symmetry number
T	temperature
t	time
τ_{P}	coupling parameter in constant pressure MD
τ_{T}	coupling parameter in constant temperature MD
V	potential energy
V	volume
V_n	torsional force constant
$\mathbf{v}$	velocity

$\mathbf{x}$	position in one dimension, unit cell vector
$\mathbf{y}$	unit cell vector
Z	total partition function
z	component of total partition function
$\mathbf{z}$	unit cell vector
$\langle\ \rangle$	ensemble average
$[\ ]$	concentration

Bibliography

[1] C. L. Brooks III, M. Karplus, and B. M. Pettitt, "Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics," *Adv. Chem. Phys.* **LXXI,** 1 (1988).

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research* **28,** 235 (2000).

[3] I. Massova and P. A. Kollman, *J. Am. Chem. Soc.* **121,** 8133 (1999).

[4] A. J. Giaccia and M. B. Kastan, *Genes Dev.* **12,** 2973 (1998).

[5] P. H. Kussie, S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine, and N. P. Pavletich, *Science* **274,** 948 (1996).

[6] A. Böttger, V. Böttger, C. Garcia-Echeverria, P. Chéne, H.-K. Hochkeppel, W. Sampson, K. Ang, S. F. Howard, S. M. Picksley, and D. P. Lane, *J. Mol. Biol.* **269,** 744 (1997).

[7] A. S. Fauci, *N. Eng. J. Med.* **341,** 1046 (1999).

[8] M. T. Schechter, K. J. Craib, K. A. Gelmon, J. S. Montaner, T. N. Le, and M. V. O'Shaughnessy, *Lancet* **341,** 658 (1993).

[9] S. C. Darby, D. W. Ewart, P. L. Giangrande, P. J. Dolin, R. J. Spooner, and C. R. Rizza, *Nature* **377,** 79 (1995).

[10] J. B. Jackson, S. Y. Kwok, J. J. Sninsky, J. S. Hopsicker, K. J. Sannerud, F. S. Rhame, K. Henry, M. Simpson, and H. H. Balfour, Jr., *J. Clin. Microbiol.* **28,** 16 (1990).

[11] "*AVERT - Evidence HIV Causes AIDS*,", AVERT, 2001, http://www.avert.org/evidence.htm.

[12] S. Jiang and A. K. Debnath, *Biochem. Biophys. Res. Comm.* **269,** 641 (2000).

[13] A. Wlodawer and J. Vondrasek, *Annu. Rev. Biophys. Biomol. Struct.* **27,** 249 (1998).

[14] R. S. Hogg, S. A. Rhone, B. Yip, C. Sherlock, B. Conway, M. T. Schechter, M. V. O'Shaughnessy, and J. S. Montaner, *AIDS* **12,** 279 (1998).

[15] F. J. Palella, K. M. Delaney, A. C. Moorman, M. O. Loveless, J. Fuhrer, G. A. Satten, D. J. Aschman, and S. D. Holmberg, *N. Eng. J. Med.* **338,** 853 (1998).

[16] A. Carr, K. Samaras, A. Thorisdottir, G. R. Kaufmann, D. J. Chisholm, and D. A. Cooper, *Lancet* **353,** 2093 (1999).

[17] D. C. Chan and P. S. Kim, *Cell* **93,** 681 (1998).

[18] D. M. Eckert and P. S. Kim, *Annu. Rev. Biochem.* **70,** 777 (2001).

[19] M. J. Root, M. S. Kay, and P. S. Kim, *Science* **291,** 884 (2001).

[20] J. G. Sodroski, *Cell* **99,** 243 (1999).

[21] D. M. Eckert, V. N. Malashkevich, L. H. Hong, P. A. Carr, and P. S. Kim, *Cell* **99,** 103 (1999).

[22] O. V. Buzko and K. M. Shokat, *Nat. Struct. Biol.* **6,** 906 (1999).

[23] M. Ferrer, T. M. Kapoor, T. Strassmaier, W. Weissenhorn, J. J. Skehel, D. Oprian, S. L. Schreiber, D. C. Wiley, and S. C. Harrison, *Nat. Struct. Biol.* **6,** 953 (1999).

[24] G. Zhou, M. Ferrer, R. Chopra, T. M. Kapoor, T. Strassmaier, W. Weissenhorn, J. J. Skehel, D. Oprian, S. L. Schreiber, S. C Harrison and D. C. Wiley, *Bioorg. Med. Chem.* **8,** 2219 (2000).

[25] S. Jiang, K. Lin, and M. Lu, *J. Virol.* **72,** 10213 (1998).

[26] R. A. Furuta, C. T. Wild, Y. Weng, and C. D. Weiss, *Nat. Struct. Biol.* **5,** 276 (1998).

[27] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley, *Nature* **387,** 426 (1997).

[28] K. Tan, J. Liu, J. Wang, S. Shen, and M. Lu, *Proc. Natl. Acad. Sci. U.S.A.* **94,** 12303 (1997).

[29] D. Fass and P. S. Kim, *Curr. Biol.* **5,** 1377 (1995).

[30] V. N. Malashkevich, B. J. Schneider, M. L. McNally, M. A. Milhollen, J. X. Pang, and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2662 (1999).

[31] V. N. Malashkevich, M. Singh, and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **98,** 8502 (2001).

[32] X. Zhao, M. Singh, V. N. Malashkevich, and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **97,** 14172 (2000).

[33] D. Fass, S. C. Harrison, and P. S. Kim, *Nat. Struct. Biol.* **3,** 465 (1996).

[34] D. R. Littman, *Cell* **93,** 677 (1998).

[35] D. C. Chan, D. Fass, J. M. Berger, and P. S. Kim, *Cell* **89,** 263 (1997).

[36] D. C. Chan, C. T. Chutkowski, and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **95,** 15613 (1998).

[37] M. Lu, S. C. Blacklow, and P. S. Kim, *Nat. Struct. Biol.* **2,** 1075 (1995).

[38] B. S. Jin, J. R. Ryu, K. Ahn, and Y. G. Yu, *AIDS* **16,** 1797 (2000).

[39] J. K. Judice, J. Y. K. Tom, W. Huang, T. Wrin, J. Vennari, C. J. Petropoulos, and R. S. McDowell, *Proc. Natl. Acad. Sci. U.S.A.* **94,** 13426 (1997).

[40] S. K. Sia, P. A. Carr, A. G. Cochran, V. N. Malashkevich, and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **99,** 14664 (2002).

[41] J. P. Moore and T. Dragic, *Nature* **401,** 759 (1999).

[42] D. M. Eckert, V. N. Malashkevich, and P. S. Kim, *J. Mol. Biol.* **284,** 859–865 (1998).

[43] T. N. M. Schumacher, L. M. Mayr, D. L. Minor, M. A. Milhollen, M. W. Burgess, and P. S. Kim, *Science* **271,** 1854 (1996).

[44] D. M. Eckert and P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **98,** 11187 (2001).

[45] C. A. Bewley, J. M. Louis, R. Ghirlando, and G. M. Clore, *J. Biol. Chem.* **277,** 14238 (2002).

[46] A. K. Debnath, L. Radigan, and S. Jiang, *J. Med. Chem.* **42,** 3203 (1999).

[47] J. E. Straub, *New Developments in Theoretical Study of Proteins* (World Scientific, Singapore, 1995).

[48] M. Karplus and E. I. Shakhnovich, *Protein Folding* (Freeman, San Francisco, 1997).

[49] P. Bamborough and F. E. Cohen, *Curr. Opin. Struct. Biol.* **6,** 236 (1996).

[50] T. E. Cheatham III and B. R. Brooks, *Theor. Chem. Acc.* **99,** 279 (1998).

[51] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science* **267,** 1619 (1995).

[52] T. Schlick, E. Barth, and M. Mandziuk, *Annu. Rev. Biophys. Biomol. Struct.* **26,** 181 (1997).

[53] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman, *Annu. Rev. Biophys. Biomol. Struct.* **30,** 211 (2001).

[54] P. A. Kollman, *Chem. Rev.* **93,** 2395 (1993).

[55] L. Salwinski and D. Eisenberg, *Curr. Opin. Struct. Biol.* **13,** 377 (2003).

[56] A. H. Elcock, D. Sept, and J. A. McCammon, *J. Phys. Chem. B* **105,** 1504 (2001).

[57] G. P. Brady and K. A. Sharp, *J. Mol. Biol.* **254,** 77 (1995).

[58] J. A. McCammon, *Curr. Opin. Struct. Biol.* **8,** 245 (1998).

[59] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, Oxford, 1989).

[60] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry* (Dover, London, 1996).

[61] W. J. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986).

[62] A. J. Stone, *The Theory of Intermolecular Forces* (Oxford University Press, Oxford, 1996).

[63] A. Hinchliffe, *Chemical Modeling From Atoms to Liquids* (Wiley, Chichester, 1999).

[64] A. Hinchliffe, *Modelling Molecular Structures* (Wiley, Chichester, 2000).

[65] J. M. Goodman, *Chemical Applications of Molecular Modelling* (Royal Society of Chemistry, Cambridge, 1998).

[66] A. R. Leach, *Molecular Modelling Principles and Applications* (Longman, Singapore, 1986).

[67] P. W. Atkins and R. S. Friedman, *Molecular Quantum Mechanics* (Oxford University Press, Oxford, 1997).

[68] M. J. Field, *A Practical Introduction to the Simulation of Molecular Systems* (Cambridge University Press, Cambridge, 1999).

[69] D. Bakowies and W. Thiel, *J. Phys. Chem.* **100,** 10580 (1996).

[70] G. Monard and K. M. Merz Jr., *Acc. Chem. Res.* **32,** 904 (1999).

[71] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, New York, 1999).

[72] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1989).

[73] P. K. Weiner and P. A. Kollman, *J. Comput. Chem.* **2,** 287 (1981).

[74] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* **117,** 5179 (1995).

[75] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkeiwicz-Kuczera, D. Yin, and M. Karplus, *J. Phys. Chem. B* **102,** 3586 (1998).

[76] C. S. Ewig, T. S. Thatcher, and A. T. Hagler, *J. Phys. Chem. B* **103,** 6998 (1999).

[77] U. Stocker and W. F. van Gunsteren, *Proteins* **40,** 145 (2000).

[78] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118,** 11225 (1996).

[79] J.-H. Lii and N. L. Allinger, *J. Comput. Chem.* **12,** 186 (1991).

[80] T. E. Cheatham III and P. A. Kollman, *Annu. Rev. Phys. Chem.* **51,** 435 (2000).

[81] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79,** 926 (1983).

[82] B. Roux and T. Simonson, *Biophys. Chem.* **78,** 1 (1999).

[83] B. Honig and A. Nicholls, *Science* **268,** 1144 (1995).

[84] K. A. Sharp and B. Honig, *J. Phys. Chem.* **94,** 7684 (1990).

[85] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.* **112,** 6127 (1990).

[86] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *J. Phys. Chem. A* **101,** 3005 (1997).

[87] M. W. Mahoney and W. L. Jorgensen, *J. Chem. Phys.* **112,** 8910 (2000).

[88] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21,** 1087 (1953).

[89] D. C. Clary, *J. Chem. Phys.* **114,** 9725 (2001).

[90] T. F. Miller III and D. C. Clary, *J. Chem. Phys.* **116,** 8262 (2002).

[91] Y.-H. Lee and B. J. Berne, *J. Phys. Chem. A* **104,** 86 (2000).

[92] Y.-H. Lee and B. J. Berne, *J. Phys. Chem. A* **105,** 459 (2001).

[93] J. W. Essex, D. L. Severance, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **101,** 9663 (1997).

[94] J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature* **267,** 585 (1977).

[95] P. Drabik, A. Liwo, C. Czaplewski, and J. Ciarkowski, *Protein Eng.* **14,** 747 (2001).

[96] R. Elber, *Curr. Opin. Struct. Biol.* **6,** 232 (1996).

[97] B. J. Berne and J. E. Straub, *Curr. Opin. Struct. Biol.* **7,** 181 (1997).

[98] P. F. Batcho, D. A. Case, and T. Schlick, *J. Chem. Phys.* **115,** 4003 (2001).

[99] P. E. Smith, *J. Chem. Phys.* **111,** 5568 (1999).

[100] W. Wriggers, E. Mehler, F. Pitici, H. Weinstein, and K. Schulten, *Biophys. J.* **74,** 1622 (1998).

[101] D. A. McQuarrie and J. D. Simon, *Molecular Thermodynamics* (University Science Books, Sausalito, California, 1999).

[102] J. M. Seddon and J. D. Gale, *Thermodynamics and Statistical Mechanics* (Royal Society of Chemistry, Cambridge, 2001).

[103] T. P. Straatsma and J. A. McCammon, *Annu. Rev. Phys. Chem.* **43,** 407 (1992).

[104] D. L. Beveridge and F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **18,** 431 (1989).

[105] W. L. Jorgensen and C. Ravimohan, *J. Am. Chem. Soc.* **83,** 3050 (1985).

[106] P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge, and P. A. Kollman, *Science* **235,** 574 (1987).

[107] P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman, *Science* **236,** 564 (1987).

[108] V. Daggett, F. Brown, and P. A. Kollman, *J. Am. Chem. Soc.* **111,** 8247 (1989).

[109] R. Zwanzig, *J. Chem. Phys.* **22,** 1420 (1954).

[110] J. G. Kirkwood, *Theory of Liquids* (Gordon and Breach, New York, 1968).

[111] J. P. Valleau and G. M. Torrie, *Modern Theoretical Chemistry* (Plenum Press, New York, 1977).

[112] J. P. M. Postma, H. J. C. Berendsen, and J. R. Haak, *Faraday Symp. Chem. Soc.* **17,** 55 (1982).

[113] A. Warshel, *J. Phys. Chem.* **86,** 2218 (1982).

[114] B. L. Tembe and J. A. McCammon, *Comput. Chem.* **8,** 281 (1984).

[115] K. M. Merz and P. A. Kollman, *J. Am. Chem. Soc.* **111,** 5649 (1989).

[116] A. Warshel, F. Sussman, and G. King, *Biochemistry* **25,** 8368 (1986).

[117] C. F. Wong and J. A. McCammon, *J. Am. Chem. Soc.* **108,** 3830 (1986).

[118] J. W. Essex, D. L. Severance, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **101,** 9663 (1997).

[119] D. M. Ferguson, R. J. Radmer, and P. A. Kollman, *J. Med. Chem.* **34,** 2654 (1991).

[120] M. R. Reddy, V. N. Viswanadhan, and J. N. Weinstein, *Proc. Natl. Acad. Sci. U.S.A.* **88,** 10287 (1991).

[121] A. Tropsha and J. Hermans, *Protein Eng.* **5,** 29 (1992).

[122] P. Cieplak and P. A. Kollman, *J. Comput. Aided. Mol. Des.* **7,** 291 (1993).

[123] B. G. Rao, R. F. Tilton, and U. C. Singh, *J. Am. Chem. Soc.* **114,** 4447 (1992).

[124] B. G. Rao and M. A. Murcko, *J. Comput. Chem.* **15,** 1241 (1994).

[125] B. G. Rao and M. A. Murcko, *Protein Eng.* **9,** 767 (1996).

[126] S. W. Rick, I. A. Topol, J. W. Erickson, and S. K. Burt, *Protein Sci.* **7,** 750 (1998).

[127] M. A. McCarrick and P. A. Kollman, *J. Comput. Aided. Mol. Des.* **13,** 109 (1999).

[128] P. Cieplak, *Mol. Sim.* **28,** 173 (2002).

[129] G. Rastelli, L. Constantino, P. Vianello, and D. Barlocco, *Tetrahedron* **54,** 9415 (1998).

[130] G. Liang, R. K. Schmidt, H. A. Yu, D. A. Cumming, and J. W. Brady, *J. Phys. Chem.* **100,** 2528 (1996).

[131] A. Pathiaseril and R. J. Woods, *J. Am. Chem. Soc.* **122,** 331 (2000).

[132] M. Zacharias, T. P. Straatsma, J. A. McCammon, and F. A. Quiocho, *Biochemistry* **32,** 7428 (1993).

[133] T. Fox, T. S. Scanlan, and P. A. Kollman, *J. Am. Chem. Soc.* **119,** 11571 (1997).

[134] T. Simonson, G. Archontis, and M. Karplus, *Acc. Chem. Res.* **35,** 430 (2002).

[135] F. T. K. Lau and M. Karplus, *J. Mol. Biol.* **236,** 1049 (1994).

[136] V. Helms and R. C. Wade, *J. Am. Chem. Soc.* **120,** 2710 (1998).

[137] H. Park and S. Lee, *J. Comput. Aided. Mol. Des.* **18,** 375 (2004).

[138] T. Jenuwein and C. D. Allis, *Science* **293,** 1074 (2001).

[139] M. Saito and A. Sarai, *Proteins* **52,** 129 (2003).

[140] P. Setny and M. Geller, *Proteins* **58,** 511 (2005).

[141] X. Kong and C. L. Brooks III, *J. Chem. Phys.* **105,** 2414 (1996).

[142] Z. Guo, C. L. Brooks III, and X. Kong, *J. Phys. Chem. B* **102,** 2032 (1998).

[143] J. Pitera and P. A. Kollman, *J. Am. Chem. Soc.* **120,** 7557 (1998).

[144] A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **63,** 1195 (1989).

[145] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.* **13,** 1011 (1992).

[146] Z. Guo and C. L. Brooks III, *J. Am. Chem. Soc.* **120,** 1920 (1998).

[147] Z. Y. Guo, J. Durkin, T. Fischmann, R. Ingram, A. Prongay, R. M. Zhang, and V. Madison, *J. Med. Chem.* **46,** 5360 (2003).

[148] M. A. L. Eriksson, J. Pitera, and P. A. Kollman, *J. Med. Chem.* **42,** 868 (1999).

[149] J. W. Pitera and P. A. Kollman, *Proteins* **41,** 385 (2000).

[150] D. L. Veenstra and P. A. Kollman, *Prot. Eng.* **10,** 789 (1997).

[151] R. J. Radmer and P. A. Kollman, *J. Comput. Aided. Mol. Des.* **12,** 215 (1998).

[152] D. A. Pearlman, *J. Med. Chem.* **42,** 4313 (1999).

[153] T. S. Lee and P. A. Kollman, *J. Am. Chem. Soc.* **122,** 4385 (2000).

[154] D. A. Pearlman and P. S. Charifson, *J. Med. Chem.* **44,** 502 (2001).

[155] D. A. Pearlman and P. S. Charifson, *J. Med. Chem.* **44,** 3417 (2001).

[156] J. Åqvist, C. Medina, and J. E. Samuelson, *Protein Eng.* **7,** 385 (1994).

[157] I. D. Wall, A. R. Leach, D. W. Salt, M. G. Ford, and J. W. Essex, *J. Med. Chem.* **42,** 5142 (1999).

[158] M. L. Lamb, J. Tirado-Rives, and W. L. Jorgensen, *Bioorg. Chem.* **7,** 851 (1999).

[159] K. B. Ljungberg, J. Marelius, D. Musil, P. Svensson, B. Norden, and J. Åqvist, *Eur. J. Pharm. Sci.* **12,** 441 (2001).

[160] J. Åqvist and J. Marelius, *Comb. Chem. High T. Scr.* **4,** 613 (2001).

[161] M. Almlöf, B. O. Brandsdal, and J. Åqvist, *J. Comput. Chem.* **25,** 1242 (2004).

[162] T. Hansson and J. Åqvist, *Protein Eng.* **8,** 1137 (1995).

[163] J. Åqvist, *J. Comput. Chem.* **17,** 1587 (1996).

[164] T. Hansson, J. Marelius, and J. Åqvist, *J. Comput. Aided. Mol. Des.* **12,** 27 (1998).

[165] M. D. Paulsen and R. L. Ornstein, *Protein Eng.* **9,** 567 (1996).

[166] J. Wang, R. Dixon, and P. A. Kollman, *Proteins* **34,** 69 (1999).

[167] W. Wang, J. Wang, and P. A. Kollman, *Proteins* **34,** 395 (1999).

[168] D. K. Jones-Hertzog and W. L. Jorgensen, *J. Med. Chem.* **40,** 1539 (1997).

[169] J. G. Chen, R. X. Wang, M. Taussig, and K. N. Houk, *J. Org. Chem.* **66,** 3021 (2001).

[170] M. M. H. van Lipzig, A. M. ter Laak, A. Jongejan, N. P. E. Vermeulen, M. Wamelink, D. Geerke, and J. H. N. Meerman, *J. Med. Chem.* **47,** 1018 (2004).

[171] K. Ersmark, I. Feierberg, S. Bjelic, E. Hamelink, F. Hackett, M. J. Blackman, J. Hulten, B. Samuelsson, J. Åqvist, and A. Hallberg, *J. Med. Chem.* **47,** 110 (2004).

[172] A. M. Asi, N. A. Rahman, and A. F. Merican, *J. Molec. Graph. Model.* **22,** 249 (2004).

[173] I. Svab, A. Alexandru, G. Vitos, and M. L. Flonta, *J. Cell. Molec. Med.* **8,** 551 (2004).

[174] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case, *J. Am. Chem. Soc.* **120,** 9401 (1998).

[175] S. Huo, I. Massova, and P. A. Kollman, *J. Comput. Chem.* **23,** 15 (2002).

[176] L. T. Chong, Y. Duan, L. Wang, I. Massova, and P. A. Kollman, *Proc. Natl. Acad. Sci. U.S.A.* **96,** 14330 (1999).

[177] B. Kuhn and P. A. Kollman, *J. Am. Chem. Soc.* **122,** 3909 (2000).

[178] B. Kuhn and P. A. Kollman, *J. Med. Chem.* **43,** 3786 (2000).

[179] W. Wang and P. A. Kollman, *J. Mol. Biol.* **303,** 567 (2000).

[180] H. Santa, M. Ylisirnio, T. Hassinen, R. Laatikainen, and M. Perakyla, *Prot. Eng.* **15,** 651 (2002).

[181] N. Nordman, J. Valjakka, and M. Perakyla, *Proteins* **50,** 135 (2003).

[182] R. C. Rizzo, S. Toba, and I. D. Kuntz, *J. Med. Chem.* **47,** 3065 (2004).

[183] M. Lepsik, Z. Kriz, and Z. Havias, *Proteins* **57,** 279 (2004).

[184] D. Rinaldo, C. Vita, and M. J. Field, *J. Biomol. Struct. Dyn.* **22,** 281 (2004).

[185] E. Schrödinger, *Ann. Phys.* **79,** 489 (1926).

[186] E. Schrödinger, *Ann. Phys.* **80,** 437 (1926).

[187] N. L. Allinger, *J. Am. Chem. Soc.* **99,** 8127 (1977).

[188] N. L. Allinger, Y. H. Yuh, and J.-H. Lii, *J. Am. Chem. Soc.* **111,** 8551 (1989).

[189] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, *Comput. Phys. Commun.* **91,** 1 (1995).

[190] O. A. T. Donini and P. A. Kollman, *J. Med. Chem.* **43,** 4180 (2000).

[191] S. Y. Noskov and C. Lim, *Biophys. J.* **81,** 737 (2001).

[192] B. Jayaram, D. Sprous, M. A. Young, and D. L. Beveridge, *J. Am. Chem. Soc.* **120,** 10629 (1998).

[193] W. Wang, W. A. Lim, A. Jakalian, J. Wang, J. Wang, R. Luo, C. I. Bayly, and P. A. Kollman, *J. Am. Chem. Soc.* **123,** 3986 (2001).

[194] J. Wang, P. Morin, W. Wang, and P. A. Kollman, *J. Am. Chem. Soc.* **123,** 5221 (2001).

[195] C. M. Reyes and P. A. Kollman, *J. Mol. Biol.* **297,** 1145 (2000).

[196] R. R. Gabdoulline and R. C. Wade, *J. Phys. Chem.* **100,** 3868 (1996).

[197] L. David, R. Luo, and M. K. Gilson, *J. Comput. Chem.* **21,** 295 (2000).

[198] T. Simonson, *Curr. Opin. Struct. Biol.* **11,** 243 (2001).

[199] X. Zou, Y. Sun, and I. D. Kuntz, *J. Am. Chem. Soc.* **121,** 8033 (1999).

[200] R. B. Hermann, *J. Phys. Chem.* **76,** 2754 (1972).

[201] G. L. Amidon, S. H. Yalkowsky, S. T. Anik, and S. C. Valvani, *J. Phys. Chem.* **72,** 2239 (1975).

[202] F. Floris and J. Tomasi, *J. Comput. Chem.* **10,** 616 (1989).

[203] D. Eisenberg and A. D. McLachlan, *Nature* **319,** 199 (1986).

[204] T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **84,** 3086 (1987).

[205] M. L. Connolly, *J. Appl. Cryst.* **16,** 548 (1983).

[206] T. J. Richmond, *J. Mol. Biol.* **178,** 63 (1984).

[207] A. Jeffrey, *Mathematics for Engineers and Scientists* (Chapman & Hall, London, 1999).

[208] A. Nicholls, K. A. Sharp, and B. Honig, "*DelPhi*,", Department of Biochemistry and Molecular Biophysics, Columbia University, NY, 1990.

[209] M. Born, *Z. Phys.* **1,** 45 (1920).

[210] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1992).

[211] J. W. Ponder, "TINKER - *Software Tools for Molecular Design, Version 3.8*,", Washington University, St. Louis, 2000, http://dasher.wustl.edu/tinker.

[212] C. W. Gear, *The Numerical Integration of Ordinary Differential Equations of Various Orders* (Report ANL 7126, Argonne National Laboratory, 1966).

[213] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, NJ, 1971).

[214] L. Verlet, *Phys. Rev.* **159,** 98 (1967).

[215] R. W. Hockney, *Methods Comput. Phys.* **9,** 136 (1970).

[216] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, *J. Chem. Phys.* **76,** 637 (1982).

[217] D. Beeman, *J. Comput. Phys.* **20,** 130 (1976).

[218] A. Y. Toukmaji and J. A. Board Jr., *Comput. Phys. Commun.* **95,** 73 (1996).

[219] J. Bader and D. Chandler, *J. Phys. Chem.* **96,** 6423 (1992).

[220] H. Schreiber and O. Steinhauser, *Biochemistry* **31,** 5856 (1992).

[221] D. York, A. Wlodawer, L. Pedersen, and T. Darden, *Proc. Natl. Acad. Sci. U.S.A.* **91,** 8715 (1994).

[222] R. P. Ewald, *Ann. Phys.* **64,** 253 (1921).

[223] T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* **98,** 10089 (1993).

[224] S. de Leeuw, J. Petram, and E. Smith, Proceedings of the Royal Society of London A **373,** 27 (1980).

[225] S. Bogusz, T. E. Cheatham III, and B. R. Brooks, *J. Chem. Phys.* **108,** 7070 (1998).

[226] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, *J. Comput. Phys.* **23,** 327 (1977).

[227] J. P. Ryckaert, *Mol. Phys.* **55,** 949 (1985).

[228] W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Phys.* **40,** 1015 (1977).

[229] H. C. Andersen, *J. Comput. Phys.* **52,** 24 (1983).

[230] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *J. Chem. Phys.* **81,** 3684 (1984).

[231] J. E. Huheey, *Inorganic Chemistry* (Harper and Row, New York, 1983).

[232] G. Herzberg, *Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules* (Van Nostrand Reinhold Company, New York, 1945).

[233] K. S. Pitzer, *J. Chem. Phys.* **12,** 310 (1944).

[234] J. D. Kemp and C. J. Egan, *J. Am. Chem. Soc.* **60,** 1521 (1938).

[235] V. L. Wu and E. F. Barker, *J. Chem. Phys.* **9,** 487 (1941).

[236] J. L. Cole and V. M. Garsky, *Biochemistry* **40,** 5633 (2001).

[237] S. Wang, J. York, W. Shu, M. O. Stoller, J. H. Nunberg, and M. Lu, *Biochemistry* **41,** 7283 (2002).